CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

IMPUTAÇÃO EM CASCATA NO CONTEXTO DA INTELIGÊNCIA ARTIFICIAL CENTRADA EM DADOS

Tarsila Gomes Bello Tavares

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Jorge de Abreu Soares

Rio de Janeiro,

Janeiro de 2025

Imputação em Cascata no Contexto da Inteligência Artificial Centrada em Dados

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Tarsila Gomes Bello Tavares

Aprovada por:

Presidente, Prof. Jorge de Abreu Soares, D.Sc. (orientador)

Prof. Diego Nunes Brandão, D.Sc.

Prof. Carlos Eduardo Ribeiro de Mello, D.Sc.

Rio de Janeiro,

Janeiro de 2025

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

T231 Tavares, Tarsila Gomes Bello
Imputação em cascata no contexto da inteligência artificial centrada em dados / Tarsila Gomes Bello Tavares. — 2025.
128f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2025.
Bibliografia : f. 125-128
Orientador: Jorge de Abreu Soares

1. Mineração de dados (Computação). 2. Estruturas de dados (Computação). 3. Inteligência artificial. I. Soares, Jorge de Abreu. (Orient.). II. Título.

CDD 005.74

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

Ao meu filho, Lucca, que me ilumina cada dia com seu sorriso e pureza, inspirando-me a ser sempre a melhor versão de mim mesmo. Que este trabalho sirva de exemplo para ele da importância da dedicação e do esforço na busca pelos próprios sonhos.

Aos meus queridos pais, Marilene Gomes e Eyler Tavares, que me ensinaram o valor do trabalho duro e da perseverança.

Este trabalho é o resultado do amor, apoio e sacrifício que vocês têm dedicado a mim ao longo de toda minha vida.

AGRADECIMENTOS

Meus pais, Marilene Gomes e Eyler Tavares, receberam meu profundo agradecimento pelo suporte e incentivo constantes, fundamentais na minha formação. Agradeço especialmente ao meu filho, Lucca, minha inspiração diária.
Expresso minha gratidão ao meu orientador, Jorge Soares, por suas orientações e amizade, essenciais para meu desenvolvimento pessoal e profissional.
Aos professores Diego Brandão e Carlos Mello, pela flexibilidade e contribuições significativas ao trabalho. Agradecimento especial ao professor Diego pelo suporte no acesso a infraestrutura utilizada para os testes.
Aos professores Kele Belloze e Ronaldo Goldschmidt, pelas contribuições substanciais apresentadas no exame de qualificação e no primeiro artigo, fundamentais para a evolução deste trabalho.
Aos amigos e familiares pelo apoio nos momentos desafiadores e ao Programa de Pós-graduação em Ciência da Computação do Cefet/RJ.
A todos, meu sincero obrigado.

RESUMO

Imputação em Cascata no Contexto da Inteligência Artificial Centrada em Dados

Tarsila Gomes Bello Tavares

Orientador:
Jorge de Abreu Soares

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

À medida que o volume global de dados aumenta, é comum encontrar bases de dados com valores ausentes, o que requer a aplicação de técnicas de imputação. Tradicionalmente, estas técnicas abordam cenários univariados, tratando a ausência de valores em uma única coluna. Este estudo propõe uma abordagem de imputação em cascata, que é capaz de tratar valores ausentes em múltiplas colunas, reintegrando valores imputados na base de dados antes da imputação do atributo subsequente, permitindo sua reutilização. Adicionalmente, investigou-se a eventual melhoria da eficiência da imputação pela binarização dos dados, segundo padrões de similaridade de ausência antes da imputação, e quais algoritmos de agrupamento apresentam resultados mais interessantes para diferentes características de bases de dados. Portanto, almeja-se avaliar e comparar o desempenho de uma imputação multivariada que utiliza a abordagem em cascata com uma pré-fase de agrupamento, aplicando diferentes classes de algoritmos dessa natureza, tais como K-modes, Agglomerative Clustering, DBSCAN e a rede neural SOM. Na maioria das simulações feitas nesse trabalho, a imputação em cascata simplificada obteve melhor desempenho em relação as demais técnicas testadas

Palavras-chave:
Dados ausentes; Imputação; Imputação em Cascata; Imputação Hot-Deck

Rio de Janeiro,

Janeiro de 2025

ABSTRACT

Imputação em Cascata no Contexto da Inteligência Artificial Centrada em Dados

Tarsila Gomes Bello Tavares

Advisor:
Jorge de Abreu Soares

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

As the global volume of data increases, it is common to encounter datasets with missing values, demanding the application of imputation techniques. Traditionally, these methods address univariate scenarios, dealing with the absence of values in a single column. This study proposes a cascade imputation approach, capable of handling missing values across multiple columns, reintegrating imputed values into the database before imputing the subsequent attribute, allowing for their reuse. Additionally, the study investigated the potential improvement in imputation efficiency by binarizing data according to patterns of similarity in absence before imputation, and identifying which clustering algorithms yield the most interesting results for different dataset characteristics. Therefore, the aim is to evaluate and compare the performance of multivariate imputation using the cascade approach with a pre-clustering phase, applying various classes of algorithms, such as K-modes, Agglomerative Clustering, DBSCAN, and the SOM neural network. In most of the simulations carried out in this work, simplified cascade imputation performed better compared to the other techniques tested.

Key-words:
Missing data; Imputation; Cascade Imputation; Hot-deck Imputation

Rio de Janeiro,

Janeiro de 2025

Sumário

I Introdução **1**

I.1 Contextualização 1

I.2 Motivação 1

I.3 Objetivos 2

I.4 Organização do Texto 4

II Referencial Teórico **5**

II.1 Imputação de dados 5

II.1.1 Ausência de dados 7

II.1.2 Tipos de Ausência 7

II.1.3 Padrões de Ausência 8

II.1.4 Mecanismo de Ausência 8

II.2 Imputação Composta 11

II.3 Imputação Múltipla 12

II.4 Imputação em Cascata 13

II.4.1 Agrupamentos 14

II.4.2 Imputação 21

II.4.3 Métrica 25

II.4.4 Teste de Hipóteses 28

III Trabalhos relacionados **31**

III.1 Seleção de trabalhos 31

III.1.1 Critérios de inclusão 31

III.1.2 Critérios de exclusão 31

III.1.3 String de Busca 32

III.1.4 Resultado da Busca 32

III.2 Imputação em Cascata 34

III.2.1 Pré-processamento 34

III.2.2 Morfologia da Ausência 35

III.2.3 Critérios quanto à ordenação dos segmentos e atributos 36
III.2.4 Imputação Sequencial com Realimentação 38
III.2.5 Avaliação dos Experimentos 39
III.2.6 Limitações do Trabalho 39
III.2.7 Comparativo dos Trabalhos de Ferlin e Montiel 40

IV Metodologia 41
IV.1 Descrição da Abordagem Proposta do Trabalho 41
IV.1.1 Formalização da abordagem proposta 43
IV.2 Conjunto de Dados Utilizados 44
IV.3 Parâmetros relativos à ausência dos dados 46
IV.4 Parâmetros do k-NN 46
IV.5 Parâmetros dos Algoritmos da Morfologia da Ausência 47
IV.5.1 K-modes e Agglomerative Clustering 48
IV.5.2 DBSCAN 48
IV.5.3 Redes *Self-Organizing Map* (SOM) 49
IV.6 Resumo da metodologia a ser aplicada 50

V Análise dos Resultados 52
V.1 Descrição dos Experimentos 52
V.2 Descrição dos Dados 53
V.2.1 Wine Dataset 53
V.2.2 Pima Indians Diabetes Dataset 56
V.2.3 Boston Housing Dataset 58
V.2.4 Abalone Dataset 60
V.2.5 Forest Fire Dataset 61
V.2.6 Magic Gamma Dataset 63
V.2.7 Glass Identification Dataset 65
V.2.8 Yeast Dataset 67
V.3 Condições ambientais dos experimentos 69
V.4 Análises dos Resultados 70
V.4.1 Comparação do desempenho da Imputação em Cascata em vários Mecanismos de Ausências 70
V.4.2 Comparação do desempenho dos Critério de Ordenação 92
V.4.3 Comparação do desempenho da Imputação em Cascata de acordo com a Classe de Correlação e Outliers das Bases 100

V.4.4 Comparação do desempenho do valor de k na Imputação em Cascata 105
V.4.5 Comparação do desempenho da Imputação em Cascata usando dados normalizados 107
V.4.6 Análise da qualidade da imputação 115

VI Considerações Finais 120
VI.1 Análise Retrospectiva 120
VI.2 Visão consolidada dos resultados 120
VI.3 Trabalhos Futuros 122

Referências 125

A : Configurações utilizadas como Critério de Ordenações 129

B : Comparação das Configurações dos Critério de Ordenações e as Classes Correlação-Outlier 131

Lista de Figuras

II.1 Imputação influenciada por outras tuplas. Fonte: van Buuren [2012] 6

II.2 Padrões de ausência em conjuntos de dados onde as linhas representam os casos observados e colunas representam as variáveis. A área destacada em cinza representa os valores faltantes. Fonte: Colin-Chevalier et al. [2022] 9

II.3 Metodologia da imputação em cascata proposta por Ferlin [2008] 14

II.4 Alguns tipos de algoritmos de agrupamento. Fonte: Fahad et al. [2014] 15

II.5 Exemplo de K-means (K = 2) com a convergência dos centróides ocorrendo na nona iteração. Fonte: Manning et al. [2008] 16

II.6 Métodos utilizados nos algoritmos hierárquicos (adaptação de Ran et al. [2022]) 17

II.7 Apresenta como as redes neurais auto-organizaveis (SOM) realiza o agrupamento. Fonte: Ciaburro and Venkateswaran [2017] 20

II.8 k-vizinhos mais próximos (Fonte: Musolf et al. [2022]) 24

II.9 Visualização do erro de imputação calculado entre a diferença do valor observado (o) e o valor predito (p) (Fonte: elaboração própria) 26

III.1 Representação do Trabalho de Montiel et al. [2018]. (a) Dados originais com valores em falta faltantes em vermelho. (b) Posições atualizadas após ordenação dos atributos por contagem de valores em falta. (c-d) Iterações de imputação, repetidas até o conjunto de dados estar completo. Os valores imputados, em verde, são utilizados nas iterações seguintes do algoritmo. (Fonte: Montiel et al. [2018]) 33

III.2 Fluxo do trabalho proposto por Ferlin [2008] 35

III.3 Representação da etapa da Morfologia da Ausência (Fonte: Adaptação da proposta de Ferlin [2008]) 36

III.4 Representação da binarização 36

III.5 Fluxo de trabalho da imputação sequencial com realimentação (Fonte: Adaptação da proposta de Ferlin [2008]) 38

IV.1 Visão simplificada da abordagem da proposta do estudo 41

IV.2 Processo completo da Imputação em Cascata (Fonte: Adaptação da proposta de Ferlin [2008])) 42

IV.3 Processo de criação das classes de correlação. (a) Matriz de correlação dos atributos. (b) Os atributos com alta correlação ($\geq$ 0,5) foram substituidos por 1 e somados. Com base no valor máximo obtido, o percentual de *outliers* do conjunto de dados em relação ao total de tuplas. (Fonte: Adaptação própria) 45
IV.4 Processo de criação das classes de *outliers*. (a) Representação do boxplot contendo os Quartis e IQR. (b) O cálculo do percentual de *outliers* para cada atributo é realizado em relação ao total de tuplas. Após isso, calcula-se a média desses percentuais para obter uma medida agregada. (Fonte: Adaptação própria) 46
IV.5 Proposta da imputação em cascata de Ferlin [2008] com a aplicação de outras categorias de agrupamento 47

V.1 Boxplot dos atributos da base *Wine* 54
V.2 Matriz de correlação da base *Wine* 55
V.3 Histogramas dos atributos da base *Wine* 55
V.4 Boxplot dos atributos da base *Pima Indians Diabetes* 56
V.5 Matriz de correlação da base *Pima Indians Diabetes* 57
V.6 Histogramas dos atributos da base *Pima Indians Diabetes* 57
V.7 Boxplot dos atributos da base *Boston Housing* 59
V.8 Matriz de correlação da base *Boston Housing* 59
V.9 Histogramas dos atributos da base *Boston Housing* 59
V.10 Boxplot dos atributos da base *Abalone* 60
V.11 Matriz de correlação da base *Abalone* 61
V.12 Histogramas dos atributos da base *Abalone* 61
V.13 Boxplot dos atributos da base *Forest Fire* 62
V.14 Matriz de correlação da base *Forest Fire* 63
V.15 Histogramas dos atributos da base *Forest Fire* 63
V.16 Boxplot dos atributos da base *Magic Gamma* 64
V.17 Histogramas dos atributos da base *Magic Gamma* 64
V.18 Matriz de correlação da base *Magic Gamma* 65
V.19 Boxplot dos atributos da base *Glass Identification* 66
V.20 Histogramas dos atributos da base *Glass Identification* 66
V.21 Matriz de correlação da base *Glass Identification* 67
V.22 Boxplot dos atributos da base *Yeast* 68
V.23 Histogramas dos atributos da base *Yeast* 68
V.24 Matriz de correlação da base *Yeast* 69

V.25 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Wine* 70
V.26 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Wine* 71
V.27 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Pima* 71
V.28 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Pima* 72
V.29 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Boston* 72
V.30 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Boston* 73
V.31 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Abalone* 74
V.32 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Abalone* 74
V.33 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Forest* 75
V.34 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Forest* 75
V.35 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Gamma* 76
V.36 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Gamma* 76
V.37 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Glass* 77
V.38 Comparação dos erros míinimos de imputação (%) por método de imputação nos dados do *Glass* 77
V.39 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Yeast* 78
V.40 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Yeast* 78
V.41 Comparação do bias de correlação médio entre os dados originais de *Wine* e os dados imputados 79

V.42 Comparação do bias de correlação mínimo entre os dados originais de *Wine* e os dados imputados 80
V.43 Comparação do bias de correlação médio entre os dados originais de *Pima* e os dados imputados 80
V.44 Comparação do bias de correlação mínimo entre os dados originais de *Pima* e os dados imputados 81
V.45 Comparação do bias de correlação médio entre os dados originais de *Boston* e os dados imputados 81
V.46 Comparação do bias de correlação mínimo entre os dados originais de *Boston* e os dados imputados 82
V.47 Comparação do bias de correlação médio entre os dados originais de *Abalone* e os dados imputados 82
V.48 Comparação do bias de correlação mínimo entre os dados originais de *Abalone* e os dados imputados 83
V.49 Comparação do bias de correlação médio entre os dados originais de *Forest* e os dados imputados 83
V.50 Comparação do bias de correlação mínimo entre os dados originais de *Forest* e os dados imputados 84
V.51 Comparação do bias de correlação médio entre os dados originais de *Gamma* e os dados imputados 84
V.52 Comparação do bias de correlação mínimo entre os dados originais de *Gamma* e os dados imputados 85
V.53 Comparação do bias de correlação médio entre os dados originais de *Glass* e os dados imputados 85
V.54 Comparação do bias de correlação mínimo entre os dados originais de *Glass* e os dados imputados 86
V.55 Comparação do bias de correlação médio entre os dados originais de *Yeast* e os dados imputados 86
V.56 Comparação do bias de correlação mínimo entre os dados originais de *Yeast* e os dados imputados 87
V.57 Comparação da similaridade média entre a base original e a imputada na base *Wine* 88
V.58 Comparação da similaridade média entre a base original e a imputada na base *Pima* 88
V.59 Comparação da similaridade média entre a base original e a imputada na base *Boston* 89
V.60 Comparação da similaridade média entre a base original e a imputada na base *Abalone* 89
V.61 Comparação da similaridade média entre a base original e a imputada na base *Forest* 90

V.62 Comparação da similaridade média entre a base original e a imputada na base *Gamma* 90
V.63 Comparação da similaridade média entre a base original e a imputada na base *Glass* 91
V.64 Comparação da similaridade média entre a base original e a imputada na base *Yeast* 91
V.65 Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Wine* 92
V.66 Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Wine* 93
V.67 Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Pima* 93
V.68 Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Pima* 94
V.69 Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Boston* 94
V.70 Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Boston* 95
V.71 Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Abalone* 95
V.72 Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Abalone* 96
V.73 Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Gamma* 96
V.74 Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Gamma* 97
V.75 Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Forest* 97
V.76 Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Forest* 98
V.77 Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Glass* 98
V.78 Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Glass* 99
V.79 Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Yeast* 99
V.80 Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Yeast* 100

V.81 Comparação dos erros médios de imputação (%) por classe correlação-outliers 101
V.82 Comparação dos erros mínimos de imputação (%) por classe correlação-outliers 101
V.83 Comparação dos bias de correlação médio por classe correlação-outliers 102
V.84 Comparação dos bias de correlação mínimo por classe correlação-outliers 102
V.85 Comparação dos erros de similaridade médios por classe correlação-outliers 103
V.86 Comparação dos erros de similaridade mínimos por classe correlação-outliers 103
V.87 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Wine Normalizado* 107
V.88 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Wine Normalizado* 108
V.89 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Iris* 108
V.90 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Iris* 109
V.91 Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Iris Normalizado* 109
V.92 Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Iris Normalizado* 110
V.93 Comparação do bias de correlação médio entre os dados originais de *Wine Normalizada* e os dados imputados 110
V.94 Comparação do bias de correlação mínimo entre os dados originais de *Wine Normalizada* e os dados imputados 111
V.95 Comparação do bias de correlação médio entre os dados originais de *Iris* e os dados imputados 111
V.96 Comparação do bias de correlação mínimo entre os dados originais de *Iris* e os dados imputados 112
V.97 Comparação do bias de correlação médio entre os dados originais de *Iris Normalizada* e os dados imputados 112
V.98 Comparação do bias de correlação mínimo entre os dados originais de *Iris Normalizada* e os dados imputados 113
V.99 Comparação da similaridade média entre a base original e a imputada na base *Wine Normalizada* 114
V.100 Comparação da similaridade média entre a base original e a imputada na base *Iris* 114
V.101 Comparação da similaridade média entre a base original e a imputada na base *Iris Normalizada* 115

Lista de Tabelas

II.1 Características das métricas de (dis)similaridades. Fonte: Adaptação de Deza and Deza [2009] e Coombes et al. [2021]. 23

IV.1 Conjunto de dados e suas principais características 46

V.1 Volume de simulações por conjunto de dados e metodologias utilizadas no estudo 53
V.2 Descritivo dos atributos da base *Wine* 54
V.3 Descritivo dos atributos da base *Pima Indians Diabetes* 56
V.4 Descritivo dos atributos da base *Boston Housing* 58
V.5 Descritivo dos atributos da base *Abalone* 60
V.6 Descritivo dos atributos da base *Forest Fire* 62
V.7 Descritivo dos atributos da base *Glass Identification* 64
V.8 Descritivo dos atributos da base *Glass Identification* 66
V.9 Descritivo dos atributos da base *Yeast* 68
V.10 Comparativo dos erros de imputação, bias de correlação e erro de similaridade por mecanismos de ausência 105
V.11 Descritivo das configurações das classes correlação-outlier 106
V.12 Comparativo dos erros de imputação, bias de correlação e erro de similaridade por mecanismos de ausência 106
V.13 Comparativo da análise descritiva entre os dados originais e os imputados 118
V.14 Resumo dos testes não paramétrico de Wilcoxon 119

A.1 Descritivo das configurações utilizadas como critério de ordenações 130

B.1 Comparativo do erro médio por configurações dos critérios de ordenações e classe correlação-outlier (base e^{12}) 132

Lista de Acrônimos

Capítulo I Introdução

I.1 Contextualização

A era atual é caracterizada por um aumento substancial na quantidade de dados em diversas áreas, marcando nossa entrada na "Era do Big Data" [Wang et al., 2014]. Este crescimento no volume, velocidade e diversidade dos dados impõe desafios significativos para os profissionais em termos de eficiência e eficácia na gestão da informação [Coneglian et al., 2017]. Adicionalmente, a evolução contínua dos dados motiva as organizações a repensarem seus modelos de negócios e processos, a fim de capitalizar sobre as vantagens competitivas que uma análise de dados eficaz pode oferecer [Fontes et al., 2016].

A existência de dados ausentes em conjuntos de dados continua sendo um obstáculo significativo, afetando diretamente a qualidade das análises. Segundo Jarrahi et al. [2022], a qualidade e a completude dos dados são essenciais para apoiar sistemas de inteligência artificial (IA), especialmente com o surgimento do conceito de Inteligência Artificial centrada em Dados (*Data-Centric AI* - DCAI) [Jarrahi et al., 2022]. Consequentemente, a importância de implementar técnicas de imputação eficazes pode influenciar significativamente os resultados das análises [Larose and Larose, 2015].

A atenção dada ao problema de dados ausentes tem crescido, levando ao desenvolvimento de metodologias eficazes como a Imputação Composta, que integra várias técnicas de mineração de dados para criar novos valores a serem imputados em conjuntos de dados [Soares, 2007]. A pesquisa sobre imputação é essencial, pois a ausência de dados pode afetar negativamente o desempenho de algoritmos de mineração de dados e aprendizado de máquina, que tipicamente não tratam dados ausentes de maneira automática, exigindo um pré-processamento cuidadoso [Veroneze, 2011].

Dedicar tempo substancial à análise e ao tratamento de dados ausentes é crucial para assegurar a confiabilidade de conjuntos de dados. Nesse contexto, técnicas de imputação avançadas demonstram progressos significativos na melhoria da qualidade dos dados imputados.

I.2 Motivação

A explosão de dados gerada pela automação de processos, aliada à integração de diferentes fontes e falhas humanas, resulta em bases de dados "poluídas". Valores ausentes, imprecisos ou

incorretos comprometem a qualidade da mineração de dados, exigindo uma etapa crucial de pré-processamento: a imputação de dados.

A ausência de dados é um obstáculo significativo [Jarrahi et al., 2022], impactando diretamente a qualidade da análise. Aroyo et al. [2021] reforçam essa questão, destacando os dados como o "aspecto mais subvalorizado" da Inteligência Artificial.

Eliminar casos ou atributos incompletos, solução simplista, gera problemas maiores, como vieses na amostra e possível perda de informações e/ou conhecimento ocultos valiosos. A recuperação dessas informações é fundamental. Em pequenos conjuntos de dados, o processo manual é viável, mas em grandes volumes, soluções automatizadas são essenciais. Assim, técnicas de aprendizado de máquina estão sendo utilizadas para estimar os valores mais prováveis para os atributos ausentes.

I.3 Objetivos

A importância de desenvolver técnicas robustas para preencher dados ausentes é enfatizada pela exigência de muitos algoritmos analíticos de operarem com conjuntos de dados completos, isentos de valores faltantes.

Dentro deste cenário, a técnica de imputação em cascata, introduzida por Ferlin [2008], visa melhorar a eficácia do processo de imputação. Essa abordagem difere das convencionais ao permitir que valores imputados sejam reintegrados à base de dados de maneira sequencial antes da imputação de atributos subsequentes, ao contrário de serem adicionados ao final do processo. Esse procedimento é repetido continuamente até que todos os dados do conjunto sejam imputados.

A motivação central deste estudo reside na busca por responder à indagação: *A imputação em cascata melhora a qualidade dos dados imputados em conjuntos de dados com valores ausentes?*. Este questionamento se desdobra através das várias etapas da metodologia de imputação em cascata, dando origem a outras indagações relevantes: *"O quanto a utilização de diferentes algoritmos de agrupamento pode afetar os resultados obtidos com a imputação em cascata?"*, *"A escolha das bases de dados influencia nos resultados?"*, *"O processo de binarização é eficaz?"*, e *"O mecanismo de ausência de dados influencia na qualidade do dado imputado?"*. Essas questões são consideradas extensões da questão principal e são fundamentais para aprofundar o entendimento do impacto da imputação em cascata, conforme introduzido neste estudo.

Adicionalmente, a presente dissertação procura aperfeiçoar a metodologia de imputação de dados proposta por Ferlin [2008]. A metodologia original propôs uma abordagem híbrida que combina técnica de agrupamento e imputação sequencial com realimentação de valores. A principal inovação do trabalho de Ferlin [2008] reside na etapa de agrupamento, que precede a imputação, fundamentando-se na morfologia da ausência. Essa morfologia envolve uma segmentação inicial que utiliza o padrão espacial da ausência como base para a classificação.

Com o objetivo de responder às perguntas de pesquisa, ampliando e superando algumas limitações do estudo de Ferlin [2008], alguns cenários foram testados:: (i) diferentemente do método MCAR (*Missing Completely At Random*) exclusivamente utilizado por Ferlin [2008], neste estudo, a introdução de dados ausentes será realizada sinteticamente por meio dos mecanismos MAR (*Missing At Random*) e MNAR (*Missing Not At Random*), além do MCAR, com três percentuais distintos de ausência (10%, 20% e 30%); (ii) foram utilizados oito conjuntos de dados, quatro dos quais são frequentemente empregados em pesquisas de imputação. Todos são inicialmente completos, mas diferem em graus de correlação entre atributos e na quantidade de *outliers*, (iii) foi conduzida uma análise comparativa que envolve as quatro principais categorias de algoritmos de agrupamento: *Partitioning-based*, *Hierarchical-based*, *Density-based* e *Model-based*. Foram selecionados algoritmos representativos que se adequam ao trabalho com dados binários e/ou categorizados; (iv) desenvolveu-se uma metodologia denominada Cascata Simplificada, que representa uma adaptação do modelo original de Ferlin [2008]. Esta adaptação omite a segmentação inicial baseada no padrão de ausência, conhecida como morfologia da ausência e, por fim, (v) procedeu-se à formalização do processo de imputação em cascata, com o objetivo de tornar o método mais abstrato e aplicável em diversos contextos de pesquisa.

A imputação em cascata é uma forma específica de imputação *hot deck*, onde os valores ausentes são preenchidos usando dados observados de tuplas semelhantes dentro do mesmo conjunto. Este método iterativo reintegra valores imputados na base de dados antes de proceder com a imputação de atributos subsequentes. Tal abordagem permite que cada passo de imputação se beneficie das atualizações mais recentes.

A expansão dos mecanismos de imputação de dados ausentes, a incluir MAR (*Missing At Random*) e MNAR (*Missing Not At Random*), representa uma inovação significativa neste estudo. Ao introduzir dados ausentes de maneira sintética sob diferentes mecanismos e com variados percentuais de ausência (10%, 20% e 30%), permite-se uma avaliação mais robusta e realista das técnicas de imputação. Essa abordagem aumenta a aplicabilidade dos resultados, refletindo uma gama mais ampla de cenários reais onde os dados podem estar ausentes por razões diversas.

A escolha minuciosa dos conjuntos de dados, garantindo a diversidade dos dados, ajuda a testar a robustez das metodologias de imputação em diferentes contextos e a entender como variáveis externas, como diferentes graus de *outliers* e correlações, podem influenciar a eficácia da imputação.

A condução de uma análise comparativa envolvendo quatro categorias principais de algoritmos de agrupamento — *Partitioning-based*, *Hierarchical-based*, *Density-based* e *Model-based*— e a seleção de algoritmos representativos para trabalhar com dados binários e/ou categorizados permite uma avaliação abrangente de diferentes métodos de agrupamento no contexto da imputação. Essa análise cuidadosa fornece *insights* valiosos sobre como diferentes estratégias de agrupamento podem

influenciar os resultados da imputação, permitindo a identificação da abordagem mais eficaz para diferentes tipos de dados e estruturas de ausência.

O desenvolvimento de uma metodologia chamada "Cascata Simplificada", uma adaptação do modelo original de Ferlin [2008], introduz uma inovação metodológica ao omitir a morfologia da ausência. Esta simplificação do processo pode oferecer ganhos significativos em termos de eficiência e praticidade, tornando o método mais acessível e rápido para aplicação em cenários reais. A Cascata Simplificada busca otimizar o processo de imputação ao eliminar etapas que, embora teoricamente úteis, podem ser redundantes ou ineficazes em práticas operacionais, proporcionando uma abordagem mais direta e possivelmente mais eficaz.

Por fim, para assegurar a reprodutibilidade da metodologia proposta neste estudo, os códigos e demais arquivos essenciais para o seu desenvolvimento foram disponibilizados em um repositório no GitHub[1].

I.4 Organização do Texto

Este documento está estruturado da seguinte forma: o Capítulo 2 aborda os fundamentos teóricos essenciais para este estudo, incluindo os principais conceitos relacionados à imputação de valores ausentes. O Capítulo 3 revisa estudos pertinentes na área, conforme documentado na literatura existente. O Capítulo 4 introduz a metodologia de Imputação em Cascata, proposta por Ferlin [2008], detalhando a abordagem metodológica que será adotada nesta dissertação. O Capítulo 5 apresenta os experimentos comparativos do método proposto, juntamente com os resultados parciais obtidos de oito bases de dados de repositórios públicos. Finalmente, o Capítulo 6 apresenta as considerações finais deste estudo.

[1] https://github.com/TarsilaT/CascadeImputation.git

Capítulo II Referencial Teórico

Neste capítulo, serão explorados os conceitos-chave relacionados ao processo de imputação de dados. São apresentados os fundamentos essenciais sobre a ausência de dados, incluindo o princípio da imputação composta, imputação múltipla e imputação em cascata, além de explorar algoritmos de agrupamento relevantes à pesquisa.

II.1 Imputação de dados

Uma das principais metas de um analista de dados é assegurar que suas análises não sejam viesadas. Trabalhar com conjuntos de dados que apresentam valores ausentes pode, em diversas situações, resultar em análises distorcidas, comprometendo a precisão na representação da realidade. Nesse contexto, o analista enfrenta duas opções para contornar o problema de dados faltantes: (i) ignorar os valores ausentes, aceitando os riscos associados a essa escolha, (ii) excluir os registros que apresentem algum valor ausente em seus componentes; ou (iii) proceder com a imputação de dados faltantes, visando reduzir as adversidades oriundas dessa ausência [Schafer and Graham, 2002].

A imputação de dados é caracterizada formalmente como uma área de investigação que visa à complementação de dados faltantes utilizando métodos estatísticos e/ou técnicas de Inteligência Artificial [Gelman and Hill, 2006]. O processo de imputação envolve a substituição de dados ausentes ou inválidos por valores estimados que, embora sejam artificiais, são considerados adequados para análise.

Conforme ilustrado na Figura II.1, o processo de imputação (representado pela bola vermelha) pode ser influenciado pelos valores existentes (representados pelas bolas verdes) em outras tuplas do conjunto de dados. Esta influência ocorre tanto para o mesmo atributo quanto para atributos diferentes [van Buuren, 2012].

Por outro lado, Little and Rubin [2019] argumentam que as imputações podem ser realizadas por meio da média ou extração de valores de uma distribuição preditiva para os dados ausentes. Essa abordagem exige o desenvolvimento de uma distribuição preditiva baseada nos dados disponíveis. Os autores propõem duas metodologias fundamentais para a elaboração dessa distribuição: a modelagem explícita, que define claramente as relações entre variáveis, e a modelagem implícita, que infere essas relações indiretamente.

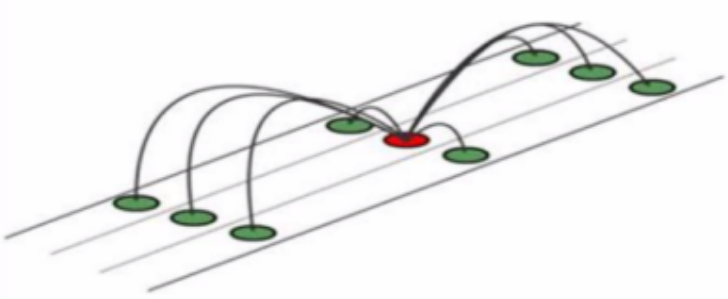


Figura II.1: Imputação influenciada por outras tuplas. Fonte: van Buuren [2012]

Na abordagem de modelagem explícita [Little and Rubin, 2019], a distribuição preditiva é fundamentada em um modelo estatístico formalmente definido, como a distribuição normal multivariada, fazendo com que as premissas do modelo sejam claramente estabelecidas. Exemplos típicos dessa abordagem incluem a imputação pela média, imputação por regressão linear e imputação por regressão estocástica.

Em contraste, a modelagem implícita concentra-se no uso de algoritmos como base do modelo, o que significa que as premissas são indiretas e necessitam de avaliação criteriosa para assegurar sua validade. Exemplos incluem a imputação *Hot Deck*, imputação por substituição e imputação *Cold Deck* [Little and Rubin, 2019].

Técnicas baseadas no *Hot Deck* preenchem valores faltantes de um atributo usando dados observados desse mesmo atributo em outros casos dentro do conjunto de dados [Schafer and Graham, 2002]. As vantagens dessas técnicas são sua simplicidade conceitual, eficácia na preservação da variabilidade dos dados e a capacidade de gerar um conjunto de dados completo, facilitando análises por métodos convencionais [Brown and Kros, 2003]. Contudo, uma limitação significativa é a complexidade em definir critérios de similaridade entre os casos, o que destaca a importância da comparação entre diferentes técnicas de agrupamento.

Por outro lado, os métodos *Cold Deck* se assemelham aos *Hot Deck*, diferindo principalmente na fonte dos dados usados para a substituição: enquanto o *Hot Deck* utiliza dados do mesmo conjunto para a imputação, o *Cold Deck* recorre a dados de um conjunto distinto para este fim [Brown and Kros, 2003].

Soares [2007] distingue o processo de imputação de dados em duas categorias principais em sua classificação: o método simples e o método híbrido. O método simples consiste na atribuição de um único valor estimado para substituir cada dado ausente no conjunto de dados. Por outro lado, o método híbrido combina duas ou mais estratégias de imputação simples com o objetivo de aprimorar a qualidade do valor estimado para os dados faltantes [Soares, 2007].

Um exemplo notável de imputação híbrida, que é central para esta pesquisa, é a imputação múltipla [Raghunathan, 2004]. Este método substitui os dados ausentes por múltiplos conjuntos de

valores plausíveis, gerando várias versões completas do conjunto de dados. Cada uma dessas versões é então analisada separadamente, utilizando estimativas estatísticas.

Soares [2007] propõe um modelo de imputação composta que investiga a influência das tarefas de pré-processamento, incluindo a seleção de atributos e o agrupamento de dados, no processo de imputação. Ferlin [2008] expande essa ideia ao introduzir a imputação em cascata. As técnicas desenvolvidas por ambos os pesquisadores serão exploradas neste capítulo.

II.1.1 Ausência de dados

O problema da falta de dados pode surgir em várias fases do ciclo de vida dos dados, abrangendo desde a coleta inicial até a transformação e integração dos dados antes de sua análise final. Com o aumento contínuo do volume de dados gerados, a ausência de dados torna-se um problema progressivamente mais significativo para analistas e pesquisadores, representando um desafio adicional tanto no âmbito acadêmico quanto no empresarial.

A importância de explorar técnicas de imputação de dados é destacada pelo fato de que a maioria dos algoritmos de análise dependerem de conjuntos de dados completos - isto é, sem a ocorrência de valores ausentes. Portanto, para a execução desses algoritmos de análise, é necessário dispor de uma base de dados sem lacunas. Nesse sentido, Rubin [1976] contribuiu significativamente para o campo ao apresentar teorias relacionadas à análise de dados faltantes, baseadas em princípios de verossimilhança, e, ao sistematizar essas abordagens, estabeleceu uma fundação teórica e experimental crucial para investigações futuras em métodos de imputação de dados.

II.1.2 Tipos de Ausência

Graham [2012] destaca na literatura a existência de dois principais tipos de ausência de dados: itens de dados ausentes (*item nonresponse*) e ondas de dados ausentes (*wave nonresponse*).

Itens de dados ausentes referem-se a situações em que, em uma pesquisa de opinião, algumas perguntas recebem respostas, enquanto outras permanecem sem resposta. Há diversas razões para que um respondente deixe de responder a uma questão específica: (i) a não observação de uma pergunta pelo entrevistado; (ii) a falta de conhecimento da resposta por parte do entrevistado; (iii) a decisão do entrevistado de não responder a uma questão por razões pessoais. Adicionalmente, a perda de respostas devido a problemas técnicos, como falhas na coleta ou no processamento dos dados, também se enquadra na categoria de itens de dados ausentes [Graham, 2012].

Por outro lado, as ondas de dados ausentes ocorrem em pesquisas realizadas em múltiplas fases ou "ondas" de coleta de dados, onde a falta de respostas se dá pela ausência completa de respostas a uma ou mais dessas fases por parte do entrevistado. Em certos casos, o entrevistado pode voltar a participar da pesquisa em momentos subsequentes para responder às questões que haviam ficado

em aberto, mas em outros casos, o entrevistado pode não retornar [Graham et al., 2012].

Além da classificação dos tipos de ausência de dados, é crucial compreender outras duas características importantes: os padrões de ausência e os mecanismos de ausência. Estas características oferecem informações adicionais sobre a natureza da ausência de dados e orientam na escolha das técnicas de imputação mais apropriadas para cada situação.

II.1.3 Padrões de Ausência

Ao analisar um conjunto de dados em busca de percepções não imediatamente aparentes, os dados são frequentemente representados em uma tabela. Dentro desse contexto, é possível identificar determinados padrões de ausência de dados. Schafer and Graham [2002] enfatizam que o reconhecimento desses padrões é um passo fundamental para selecionar a técnica de imputação mais adequada. Eles categorizam os padrões de ausência em duas grandes classes: (i) geral ou aleatório, e; (ii) específico.

No caso de ausências de padrão geral ou aleatório, como o nome sugere, os dados faltantes estão distribuídos de maneira dispersa pelo conjunto de dados, sem um padrão discernível de onde as ausências ocorrem. Isso implica que qualquer registro pode conter dados ausentes, independentemente de outras variáveis.

Quanto aos padrões de ausência específicos (Figura II.2), estes são divididos em dois subtipos: (i) univariados, e; (ii) monotônicos. Os padrões específicos classificados como univariados caracterizam-se pela concentração de dados ausentes em uma única variável. Ou seja, a ausência ocorre especificamente em um item Y, enquanto o restante p dos itens $X_1, \ldots, X_p$ permanece completo. Este padrão também inclui situações em que Y representa um conjunto de itens que estão ou totalmente presentes ou totalmente ausentes, essencialmente restritos a uma variável específica do conjunto de dados.

Por outro lado, nos padrões monotônicos, as ausências ocorrem em mais de uma variável de tal forma que, se um conjunto de itens $Y_1, \ldots, Y_p$ é organizado em uma certa ordem e se Y_j possui valores ausentes, então itens $Y_j+1, \ldots, p$ também terão dados ausentes (Figura II.2). Este padrão sugere uma relação sequencial entre as variáveis, onde a presença de dados ausentes em uma implica a ausência em variáveis subsequentes [Schafer and Graham, 2002].

II.1.4 Mecanismo de Ausência

Ao abordar a imputação de dados, é essencial não apenas identificar os padrões de ausência no conjunto de dados, mas também compreender os mecanismos subjacentes que levam a essa ausência. O mecanismo de ausência refere-se às causas subjacentes que resultam na não geração de alguns dados. Esta compreensão é crucial para escolher a estratégia de imputação mais adequada e para

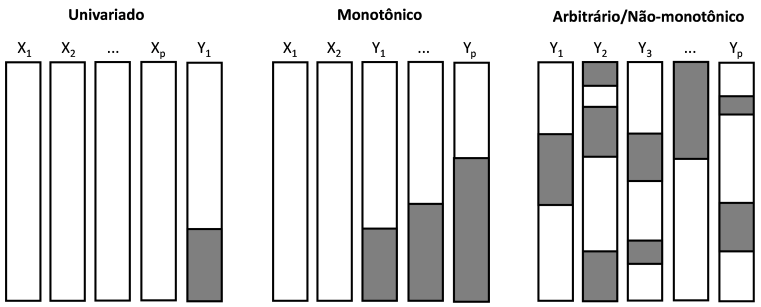


Figura II.2: Padrões de ausência em conjuntos de dados onde as linhas representam os casos observados e colunas representam as variáveis. A área destacada em cinza representa os valores faltantes. Fonte: Colin-Chevalier et al. [2022]

interpretar corretamente os resultados das análises.

De acordo com a literatura estatística, os mecanismos de ausência descrevem o processo através do qual os dados são ou não gerados [Schafer and Graham, 2002]. Little and Rubin [2019] enfatizam que entender o mecanismo de ausência é fundamental, pois ele está relacionado ao processo que leva à não geração dos dados. Eles também ressaltam a importância de conhecer a relação entre os dados ausentes e os valores reais das variáveis envolvidas no conjunto de dados.

Little and Rubin [2019] propõem que se Y representa um conjunto de dados completo e M o conjunto indicando a presença ou ausência de dados, o mecanismo de ausência é caracterizado pela distribuição condicional de M dado Y, expresso como f(M|Y,θ), onde θ representa parâmetros desconhecidos.

Rubin [1976] introduz três classificações principais para os mecanismos de ausência de dados, cada um implicando diferentes suposições sobre a natureza dos dados ausentes e estratégias adequadas para lidar com eles:

- **Completamente Aleatório (MCAR – *Missing Completely At Random)***: este mecanismo assume que a ausência de dados é totalmente aleatória, sem qualquer correlação com outros atributos do conjunto de dados. Em outras palavras, a probabilidade de um dado estar ausente é a mesma para todas as observações, independente dos valores observados ou não observados em Y. Formalmente, se a distribuição condicional de M dado Y, f(M|Y,θ), é igual à distribuição de M, f(M|θ), para todos os valores de Y e θ, então o mecanismo de ausência é considerado completamente aleatório. Isso implica que a ausência de dados não introduz viés na análise, pois não está relacionada a qualquer característica dos dados.

- **Aleatório (MAR – *Missing At Random)***: neste caso, a ausência de dados está relacionada aos dados observados, mas não aos dados ausentes em si. Isso significa que, mesmo que os dados ausentes tenham uma probabilidade diferente de ocorrência em diferentes partes do conjunto de dados, essa probabilidade só depende dos valores observados. Seja Y_{obs} um

subconjunto de Y que contém todos os dados observados e Y_{mis} um subconjunto de Y que contém os dados ausentes. Para que a ausência seja classificada como aleatória, a distribuição condicional de M dado Y, f(M|Y,θ), deve ser igual à distribuição condicional de M dado apenas os dados observados, f(M|Y_{obs},θ), para todos os valores de Y_{mis} e θ.

- **Não Aleatório (NMAR – *Not Missing At Random)***: este mecanismo ocorre quando a probabilidade de um dado estar ausente está diretamente relacionada aos valores ausentes. Em outras palavras, a ausência dos dados depende das informações que não foram observadas. Esse mecanismo é considerado inacessível, segundo Graham [1993], porque a razão da ausência dos dados não pode ser medida diretamente, tornando os dados ausentes indisponíveis para análise ou avaliação. Lidar com dados NMAR requer métodos específicos que assumam ou modelam a relação entre a probabilidade de ausência e os dados ausentes, uma vez que ignorar essa dependência pode levar a análises enviesadas.

Após a identificação dos mecanismos que levam à ausência de dados, existem várias estratégias para mitigar esse problema, incluindo a possibilidade de substituição dos dados faltantes por valores reais [Rubin, 1988]. Contudo, a capacidade de muitos softwares estatísticos para tratar a ausência de dados é limitada, frequentemente restringindo-se a algumas opções principais: (i) a remoção completa de casos onde falta qualquer dado (*Listwise Deletion* ou *Complete-Case Deletion*), (ii) a remoção de pares de dados onde só se excluem as observações ausentes em análises específicas (*Pairwise Deletion*), e (iii) a eliminação de colunas inteiras do conjunto de dados que contêm valores ausentes. Uma alternativa adicional adotada por esses softwares é a substituição dos dados ausentes por alguma medida de resumo, como a média ou a mediana [Assunção, 2012].

No entanto, Ferlin [2008] ressalta que o uso da remoção completa de casos pode levar a diversos problemas na análise dos dados. Esses problemas incluem a redução significativa do tamanho da amostra, o que pode afetar a precisão dos resultados; o descarte de uma grande quantidade de informações, introduzindo potenciais vieses; e, no contexto de classificação ou sumarização, a eliminação de regras importantes. De forma semelhante, a remoção de pares apresenta desafios, como a dificuldade em realizar comparações consistentes dentro de um mesmo estudo, devido ao uso de diferentes subconjuntos de casos, e a possibilidade de que a matriz de covariância não possa ser adequadamente definida, pois certos elementos podem se tornar indefinidos em função de outros.

Com os desafios impostos pelas limitações das técnicas convencionais de tratamento de dados ausentes, pesquisadores têm encontrado formas de aperfeiçoar os métodos de imputação de dados. Uma abordagem promissora envolve a utilização de workflows científicos para automatizar não apenas a configuração, mas também a execução e a avaliação das técnicas de imputação [Castañeda et al., 2008]. Essa inovação representa um avanço significativo na área, pois permite uma otimização

do processo de imputação, garantindo uma aplicação mais eficiente e eficaz das técnicas disponíveis. A próxima seção deste trabalho se dedicará a discutir em detalhe essas melhorias, explorando como a automação e a implementação de estratégias de imputação mais sofisticadas podem contribuir para a qualidade dos conjuntos de dados tratados e, consequentemente, para a precisão das análises realizadas com base neles.

II.2 Imputação Composta

A técnica de imputação composta [Soares, 2007] representa uma abordagem estratégica e planejada para a imputação de dados ausentes. Esta metodologia se apoia na premissa de que um planejamento cuidadoso, incorporando tarefas de seleção, agrupamento ou classificação antes da imputação, pode significativamente melhorar a qualidade dos dados imputados. Essa abordagem permite uma preparação mais adequada do conjunto de dados, visando uma imputação mais precisa e eficaz.

O processo de imputação composta [Soares, 2007] é estruturado nas seguintes etapas:

- uma sequência de tarefas T_i: seleção de atributos, agrupamento ou classificação (como, T_1 = seleção de atributos, T_2 = agrupamento, T_3 = criação de regras de associação, T_4 = imputação, entre outros).

- $\rightarrow$: operador que define uma ordem de precedência. Por exemplo, a expressão $X \rightarrow Y$ indica que a tarefa X ocorre antes da tarefa Y.

- E(ν,B): estratégia de imputação para um dado atributo ν de um conjunto de dados B é representada por uma sequência de tarefas $T_1 \rightarrow T_2 \rightarrow ... \rightarrow T_m$, onde T_m é sempre uma tarefa de imputação

- A_i: algoritmo utilizado no processo. Para executar essas tarefas, diversos algoritmos podem ser utilizados, por exemplo, a média ou o algoritmo dos k vizinhos mais próximos.

- $\Rightarrow$: indica a ordem em que esses algoritmos são aplicados, onde $A_i \Rightarrow A_j$ significa que o algoritmo A_i é aplicado antes do A_j.

- P(ν,B): plano de imputação de um atributo ν de um conjunto de dados B é formulado como uma sequência ordenada de algoritmos $A_1 \Rightarrow A_2 \Rightarrow ... \Rightarrow A_p$, onde A_p é necessariamente um algoritmo de imputação.

- Ψ: instância de aplicação do algoritmo A_i, que é definida de acordo com um conjunto de parâmetros $\boldsymbol{\theta}_i = \{\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,q}\}, \quad \Psi = f(A_i, \boldsymbol{\theta}_i)$.

- I(ν,B) : instância de um plano de imputação de um atributo ν de um conjunto de dados B, composta por uma sequência ordenada de q instâncias de aplicação de algoritmos. Logo, $\Psi_1 \Rightarrow \Psi_2 \Rightarrow \cdots \Rightarrow \Psi_q$, onde Ψ_q é, obrigatoriamente, uma instância de aplicação de um algoritmo de imputação.

- ε(I(ν,B)): uma medida de erro que avalia a eficácia do plano de imputação do atributo ν.

O objetivo final é selecionar o conjunto de valores gerados pelo plano de imputação que apresente o menor erro médio em comparação a todas as instâncias avaliadas desse plano, garantindo assim a melhor qualidade de dados imputados possível. Esse método permite uma abordagem mais refinada e adaptativa à imputação, destacando-se pela sua capacidade de incorporar conhecimento específico do domínio e pela flexibilidade na escolha e na aplicação de técnicas de imputação.

II.3 Imputação Múltipla

No método de imputação simples, também conhecido como imputação única, gera-se apenas um valor possível para cada campo ausente no conjunto de dados. Apesar de sua simplicidade, esse método pode estimar valores com uma variância reduzida, o que pode resultar em análises enviesadas, conforme apontado por Cartwright et al. [2003].

A imputação simples pode ser realizada através de vários métodos estatísticos, incluindo a média, a moda, a regressão simples e a regressão logística. Contudo, uma limitação importante dessa abordagem é que ela tende a ocultar a incerteza dos dados imputados, invalidando potencialmente os intervalos de confiança por basear-se exclusivamente nos dados existentes.

Para superar esses desafios, a imputação múltipla surge como uma solução avançada. Definida por Rubin [1988], a imputação múltipla cria vários conjuntos de dados completos, imputando valores distintos para refletir a incerteza sobre o valor correto a ser imputado. Após a imputação, esses conjuntos de dados são analisados utilizando procedimentos padrões de análise, e os resultados de cada análise são então combinados para produzir o resultado final. Essencialmente, essa abordagem estima "n" valores sugeridos para cada atributo com valores ausentes, preenchendo-os em "n" conjuntos de dados simulados, como se fossem realizadas "n" imputações simples. Esses conjuntos são posteriormente avaliados, considerando a variância entre eles, a fim de produzir uma imputação consolidada e mais precisa.

A imputação múltipla, portanto, leva em consideração a incerteza associada aos valores estimados, possibilitando a produção de estimativas de atributos não enviesadas. Segundo Wayman [2003], esse método oferece uma maneira mais robusta de tratar dados ausentes, gerando estimativas mais confiáveis. Rubin [1988] descreve a imputação múltipla como uma extensão da imputação simples, onde cada atributo ausente é substituído por um conjunto de valores estimados, diminuindo signifi-

cativamente a incerteza inerente ao processo de imputação e melhorando a qualidade da análise de dados.

II.4 Imputação em Cascata

Utilizando as vantagens da imputação múltipla, foi desenvolvida a técnica de imputação em cascata como resposta ao desafio de atribuir valores em conjuntos de dados que contêm diversos atributos numéricos faltantes. Essa metodologia, proposta inicialmente por Ferlin [2008], mostra-se particularmente vantajosa em cenários onde vários atributos de um mesmo registro faltam simultaneamente.

O procedimento de imputação em cascata começa com uma etapa de agrupamento, onde os registros incompletos são organizados em grupos com base na morfologia dos dados ausentes que apresentam. A morfologia em questão diz respeito ao padrão espacial que os dados ausentes formam nas tuplas, levando em consideração a distribuição dos dados presentes e ausentes [Ferlin, 2008]. Esse exame detalhado das posições relativas dos dados faltantes auxilia na identificação de conjuntos de casos que compartilham características de ausência semelhantes.

Nessa etapa inicial, ocorre a binarização dos dados dos registros incompletos, convertendo os valores em uma representação binária onde os dados faltantes são indicados por 1 e os presentes por 0. Esse método gera um conjunto de dados formado exclusivamente por 0s e 1s, refletindo de maneira direta quais informações estão ausentes ou presentes em cada atributo dos registros.

Com essa binarização, o agrupamento se torna mais eficaz, uma vez que é possível comparar os registros baseando-se na semelhança dos padrões de ausência, independentemente dos valores específicos dos dados. Portanto, essa estratégia de pré-processamento viabiliza uma compreensão objetiva da estrutura de ausência no conjunto de dados, direcionando a criação de grupos coesos para a subsequente imputação em cascata. Este método cria um fundamento robusto para a implementação de algoritmos de imputação subsequentes, promovendo um tratamento dos dados ausentes de maneira precisa e fundamentada.

Seguindo a organização dos grupos conforme a morfologia de ausência , a imputação é executada sequencialmente de um grupo para outro. Os dados imputados em cada fase são utilizados para apoiar a imputação das fases seguintes, provocando um efeito cascata (Figura II.3). Esse processo não apenas melhora a precisão da imputação ao considerar a dinâmica entre os dados ausentes e presentes, mas também facilita uma maneira mais ordenada e eficaz de abordar a complexidade dos dados faltantes em múltiplos atributos [Ferlin, 2008].

A estratégia de imputação em cascata é, portanto, um marco importante para o manejo de dados ausentes, especialmente em bases de dados volumosas e complexas, onde técnicas convencionais de imputação podem não ser adequadas para recuperar a integridade dos dados de forma satisfatória.

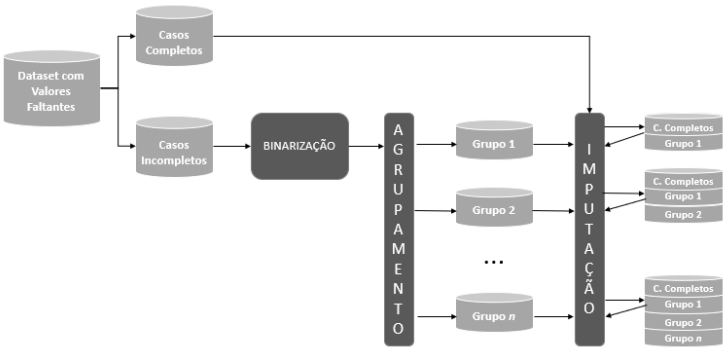


Figura II.3: Metodologia da imputação em cascata proposta por Ferlin [2008]

Adicionalmente, a variação na morfologia de ausência possibilita a flexibilidade na escolha dos métodos de agrupamento ou imputação. Isso significa que cada conjunto de casos pode ser tratado com uma técnica de imputação específica, permitindo que grupos com padrões de ausência menos complexos sejam processados por métodos mais simples, enquanto aqueles com padrões mais elaborados requerem técnicas de maior capacidade computacional, como discutido por Castañeda et al. [2008].

Na seção II.4.1 será detalhado as possibilidades de algoritmos de agrupamento que podem ser utilizados na identificação da morfologia dos valores ausentes. Posteriormente, na seção II.4.2, apresentaremos as alternativas mais usuais de imputação de dados.

II.4.1 Agrupamentos

Na pesquisa sobre métodos de imputação de dados, nota-se que grande parte dos estudos concentra-se na imputação simples. Soares [2007] apontou para essa tendência e, em sua pesquisa, introduziu a abordagem denominada imputação composta. Esse método visa principalmente à combinação de uma ou mais tarefas antes da imputação dos dados, incluindo, por exemplo, o uso de algoritmos de aprendizado de máquina para agrupamento de dados.

Os algoritmos de agrupamento têm-se destacado como uma ferramenta eficaz para o tratamento do grande volume de dados produzidos por aplicações modernas. Seu principal objetivo é classificar os dados em grupos, ou *clusters*, agrupando objetos similares em um mesmo segmento conforme métricas específicas [Bano and Khan, 2018].

Fahad et al. [2014] realizaram uma análise comparativa entre vários algoritmos de agrupamento, tanto do ponto de vista teórico quanto empírico. Eles propuseram uma estrutura de categorização baseada nas principais características desses algoritmos, com o intuito de estabelecer uma taxonomia para eles. Em um estudo similar, feito por Bano and Khan [2018], também categorizaram os algoritmos de agrupamento, seguindo uma abordagem parecida com a de Fahad et al. [2014].

Os métodos utilizados nos diversos algoritmos de agrupamento podem ser amplamente divididos

nas seguintes categorias: (i) *Partitioning-based*; (ii) *Hierarchical-based*; (iii) *Density-based*; (iv) *Grid-based* e (v) *Model-based*. Uma visão geral dessas categorias e dos algoritmos associados a cada uma é apresentada na Figura II.4, detalhando-se mais a fundo nas seções subsequentes do estudo. Essa organização facilita a compreensão das diferentes abordagens de agrupamento e suas aplicações potenciais, especialmente quando integradas em estratégias de imputação de dados como a imputação composta proposta por Soares [2007].

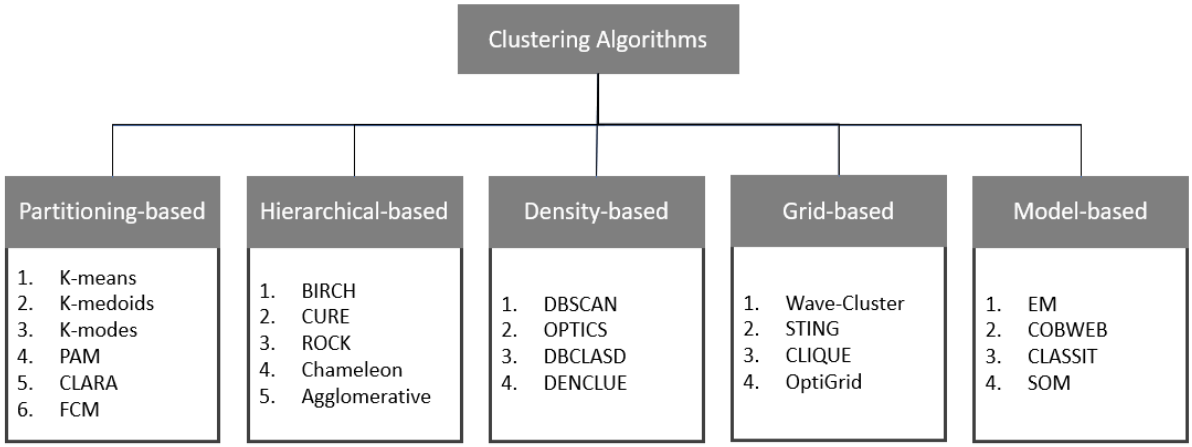


Figura II.4: Alguns tipos de algoritmos de agrupamento. Fonte: Fahad et al. [2014]

Ao selecionar os algoritmos para este estudo, considerou-se essencialmente a existência de bibliotecas em Python já implementadas e reconhecidas na área de Computação, como Scikit-learn[2], para minimizar possíveis problemas no futuro. Além disso, deu-se preferência a algoritmos de agrupamento capazes de processar dados categóricos/binários ou que permitissem a alteração do tipo de distância utilizada afim de aceitar esse tipo de dados.

Algoritmos do tipo *Partitioning-based*

Os algoritmos baseados em partição atuam subdividindo os objetos de um conjunto de dados em várias partes, cada uma representando um grupo distinto, conforme explicado por Fahad et al. [2014]. Esses algoritmos assumem que o centro geométrico ou ponto médio dos dados dentro de um conjunto reflete o núcleo do respectivo grupo.

A Figura II.5 demonstra como a técnica se baseia no cálculo iterativo dos centróides, que são os pontos centrais ou médias de um conjunto de dados, para agrupar os dados mediante a minimização da distância dos objetos a esses centróides, como descrito por Jain [2008]. Entre os algoritmos mais reconhecidos dessa categoria, encontram-se o K-means e o K-medoids, conforme discutido por Xu and Tian [2015].

Para esta categoria, escolheu-se o algoritmo K-modes, por ser próprio para dados categóricos e pela simplicidade e rapidez na verificação da (dis)similaridade entre os pontos de dados. Como o

[2]https://scikit-learn.org/stable/index.html

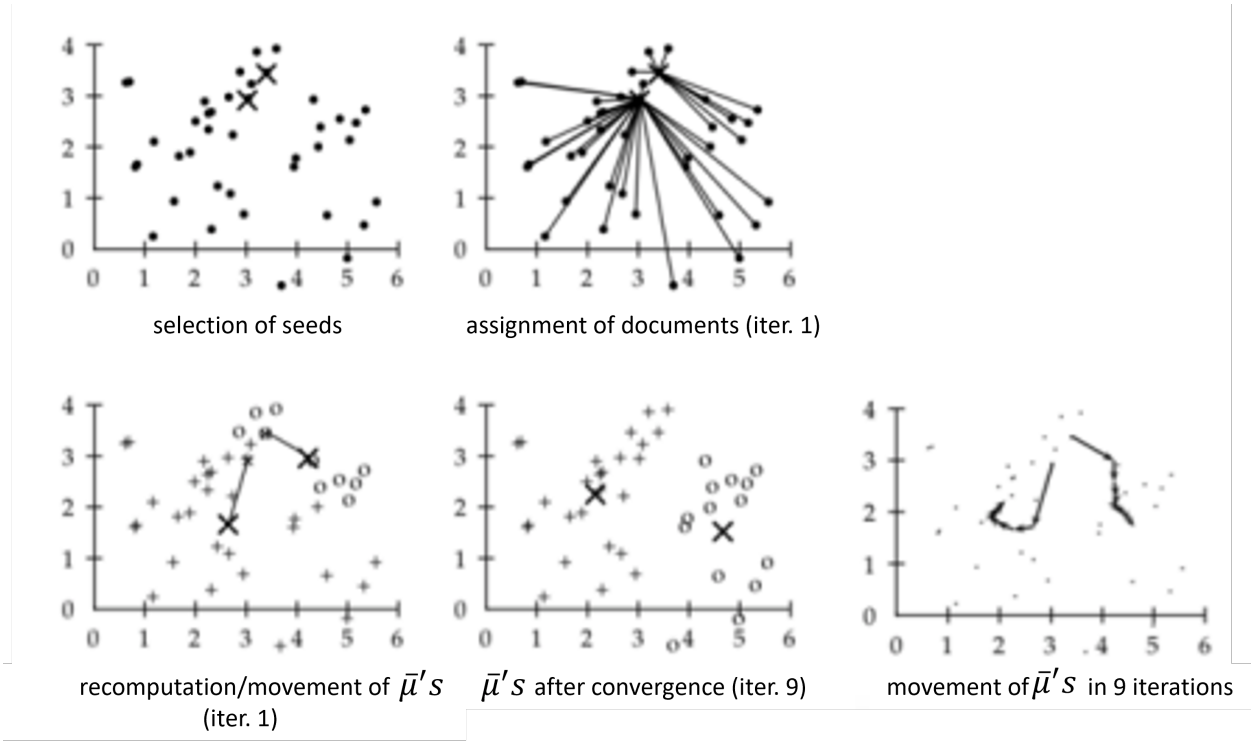


Figura II.5: Exemplo de K-means (K = 2) com a convergência dos centróides ocorrendo na nona iteração. Fonte: Manning et al. [2008]

K-modes é uma extensão do K-means, esse algoritmo tem como vantagens a sua escalabilidade. O pseudocódico do K-modes é representado na Figura 1.

Algoritmo 1 Algoritmo de Agrupamento K-Modes

Require: Dataset (D), K, max_iterations
Ensure: Labels dos clusters para todos os pontos de D
1: **Begin**: Seleciona K modas iniciais aleatoriamente do Dataset.
2: **repeat**
3: Atribuir cada ponto ao cluster mais próximo com base numa medida de similariedade (exemplo: distância de Hamming).
4: Atualizar a moda de cada cluster para a categoria mais frequente de atributos no cluster.
5: **until** Convergência ou atingir o max_iterations
6: **return** Clusters e suas modas.

Algoritmos do tipo *Hierarchical-based*

Em algoritmos baseados em hierarquia, a similaridade entre os objetos é estruturada em uma forma hierárquica, isto é, as relações entre objetos (ou grupos de objetos) são representadas através de uma árvore, frequentemente denominada dendrograma. Nessa representação, objetos que compartilham maior semelhança são posicionados em ramos próximos dentro da árvore. O comprimento desses ramos indica o nível de similaridade, documentando a sequência de fusões e divisões ocorridas durante o processo de agrupamento. De acordo com Fahad et al. [2014], a árvore gerada por esses métodos pode ser convertida em partições por meio de um corte no dendrograma em um

determinado nível.

Quanto à construção da árvore hierárquica, os algoritmos hierárquicos se dividem em dois principais tipos: divisivos (de cima para baixo) e aglomerativos (de baixo para cima), como mostra a Figura II.6. Nos algoritmos divisivos, o processo inicia com um único conjunto agregado de objetos que, a cada etapa, é dividido em grupos menores provenientes da iteração anterior, continuando essa divisão até que se formem n grupos distintos. Por outro lado, os algoritmos aglomerativos começam com n grupos individuais, considerando cada objeto da coleção como um grupo separado, e, de maneira iterativa, combinam esses grupos menores em conjuntos progressivamente maiores, até que se forme um único grupo abrangendo todos os objetos, como explicado por Bano and Khan [2018].

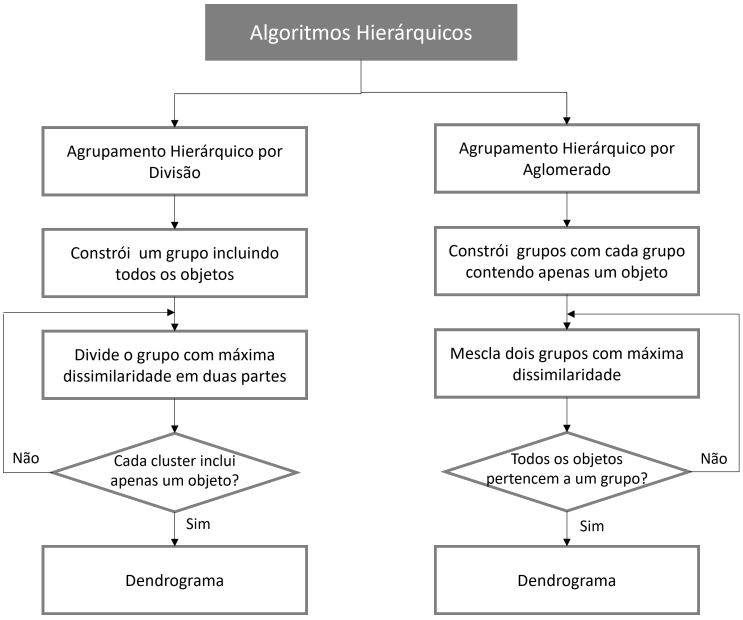


Figura II.6: Métodos utilizados nos algoritmos hierárquicos (adaptação de Ran et al. [2022])

Na literatura especializada, encontram-se diversos algoritmos aglomerativos, os quais são classificados conforme o critério adotado para calcular a proximidade entre os grupos que serão unidos. Segundo Yim and Ack Baraly [2015], as principais categorias desses critérios são: *single-linkage*, *average-linkage* e *complete-linkage*.

A categoria *single-linkage* adota uma perspectiva otimista para determinar a distância entre dois grupos, baseando-se na menor distância existente entre quaisquer objetos pertencentes a cada um dos grupos em análise. Essa abordagem permite a identificação de cadeias de objetos próximos, mesmo que estendam-se por regiões do espaço de dados [Yim and Ack Baraly, 2015].

Por outro lado, segundo Yim and Ack Baraly [2015], o *complete-linkage* emprega uma visão mais conservadora ou pessimista, onde a distância entre dois grupos é definida pela maior distância entre seus componentes. Esse método tende a produzir grupos mais compactos e bem definidos, minimizando a possibilidade de unir grupos distantes.

No *average-linkage*, a distância entre dois grupos é calculada com base na média das distâncias entre todos os pares de objetos dos dois grupos. Essa abordagem equilibra as perspectivas otimista e pessimista, levando a uma agregação que considera a disposição geral dos objetos dentro dos grupos [Yim and Ack Baraly, 2015].

Nesta categoria optou-se pelo algoritmo *Agglomerative*, dado que seu código oferece a flexibilidade de alterar o tipo de distância utilizada, permitindo a adoção da distância Gower. Além disso, o critério utilizado para calcular a proximidade entre os grupos foi o *complete-linkage*. O pseudocódigo do *Agglomerative Cluster* está representado na Figura 2.

Algoritmo 2 Algoritmo Hierarquico Aglomerativo

Require: Dataset (D), número de clusters (K)
Ensure: Labels dos clusters para todos os pontos do D
1: **Begin**: Cada ponto como um único cluster
2: Calcula a matriz de distâncias para todos os clusters
3: **while** o número de clusters $> K$ **do**
4: Encontrar os dois clusters mais próximos com base na matriz de distâncias
5: Fundir estes dois clusters num só
6: Atualizar a matriz de distância para refletir esta fusão
7: **end while**
8: **return** O conjunto final de clusters

Algoritmos do tipo *Density-based*

Os algoritmos baseados em densidade distinguem objetos com base em suas regiões de densidade, conectividade e limite, sendo essencialmente influenciados pela proximidade dos vizinhos mais próximos de um ponto [Bano and Khan, 2018]. Um cluster é identificado como um componente densamente conectado e pode expandir-se em qualquer direção guiada pela densidade local. Essa característica permite que tais algoritmos identifiquem grupos com contornos arbitrários e ofereçam proteção natural contra ruídos ou outliers [Xu and Tian, 2015]. Algoritmos como DBSCAN, OPTICS, DBCLASD e DENCLUE exemplificam essa abordagem, capacitando-os a filtrar ruídos e descobrir agrupamentos de formas variadas [Fahad et al., 2014].

Introduzido por Ester et al. [1996], o algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) é um método não paramétrico que se concentra na quantidade de pontos dentro de um raio específico. Essa abordagem permite a identificação de clusters baseada diretamente nas características dos dados, eliminando a necessidade de especificar a priori o número de grupos, o que facilita a formação de agrupamentos com formatos arbitrários [Ester et al., 1996]. Este algoritmo é adequado para conjuntos de dados volumosos que incluem outliers, sendo capaz de identificar clusters de diversos formatos.

No entanto, o DBSCAN requer a definição de dois parâmetros iniciais: (i) ϵ: é o raio de vizi-

nhança que estabelece a área de proximidade para cada ponto no conjunto de dados, permitindo ao algoritmo verificar a densidade de pontos dentro desse raio e formar um cluster se o número de pontos exceder um limiar específico; e (ii) η: é o número mínimo de pontos que define a quantidade mínima de vizinhos que um ponto deve ter dentro do raio estipulado para ser considerado um ponto central, propiciando, assim, a formação de um novo cluster conforme os critérios de agrupamento baseado em densidade. O pseudocódico do DBSCAN está representado na Figura 3.

Algoritmo 3 Algoritmo de Agrupamento DBSCAN

Require: Dataset (D), ϵ, $MinPts$
Ensure: Labels dos clusters para todos os pontos de D
  for Para cada ponto não visitado (P) em D **do**
    if |RangeQuery(D, distFunc, P, ϵ)| $< MinPts$ **then**
      label(P) $\leftarrow$ Ruído **Continue**
    end if
    $C \leftarrow C + 1$
    label(P) $\leftarrow C$
    ExpandCluster(D, distFunc, $P, \epsilon, MinPts, C$)
  end for
  return C

  procedure ExpandCluster(Dataset (D), função de distância ($distFunc$), ponto (P), ϵ, $MinPts$, C)
    for Para cada ponto Q em RangeQuery(D, distFunc, P, ϵ): **do**
      if label(Q) $\in$ {indefinido, Ruído} **then**
        label(Q) $\leftarrow C$
        if |RangeQuery(D, distFunc, Q, ϵ)| $\geq MinPts$ **then**
          ExpandCluster(D, distFunc, $Q, \epsilon, MinPts, C$)
        end if
      end if
    end for
  end procedure

Algoritmos do tipo *Model-based*

Os algoritmos baseados em modelo (*Model-based*) buscam aperfeiçoar a concordância entre os dados observados e determinados modelos matemáticos pré-estabelecidos, fundamentando-se na premissa de que os conjuntos de dados são originados a partir de uma combinação de distribuições de probabilidade. Tais algoritmos destacam-se pela capacidade de identificar automaticamente clusters, levando em conta padrões estatísticos comuns e a existência de outliers, proporcionando uma abordagem robusta para o agrupamento de dados [Bano and Khan, 2018].

Dentro do espectro dos algoritmos *model-based*, existem duas vertentes principais: as abordagens estatísticas e as que utilizam redes neurais. Um exemplo proeminente da abordagem estatística é o algoritmo *Expectation Maximization* (EM), que opera com base em um modelo de mistura de densidades. Por outro lado, no contexto das redes neurais, as redes neurais auto-organizáveis

(*Self-Organizing Map* - SOM ou Mapas de Kohonen) exemplificam a aplicação dessa tecnologia no agrupamento baseado em modelo [Bano and Khan, 2018].

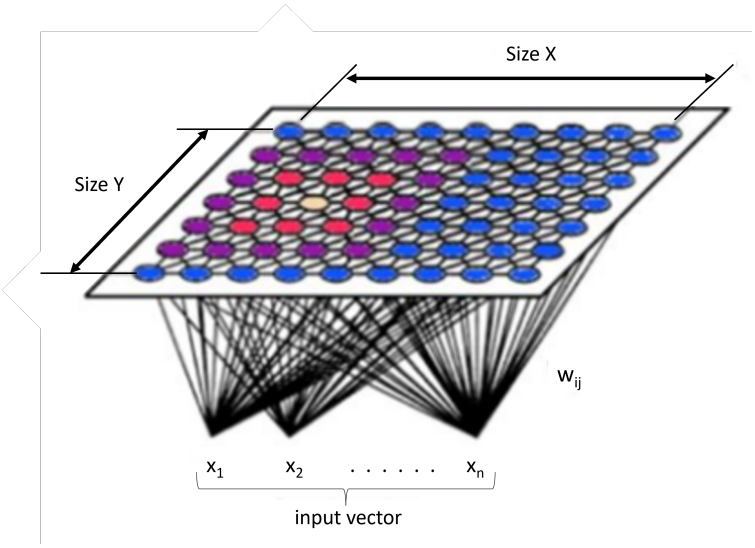


Figura II.7: Apresenta como as redes neurais auto-organizaveis (SOM) realiza o agrupamento. Fonte: Ciaburro and Venkateswaran [2017]

O processo de agrupamento em um SOM ocorre por meio de uma representação bidimensional, conforme mostra a Figura II.7, onde cada neurônio (ou célula) no mapa corresponde a um cluster potencial. Durante o treinamento, os padrões de entrada são apresentados à rede, e os pesos dos neurônios são ajustados de modo a refletir as características dos dados [Fausett, 1994]. Além do neurônio vencedor, os pesos dos neurônios vizinhos também são ajustados, mas em menor grau, preservando a topologia dos dados de entrada. Este mecanismo de ajuste não apenas classifica os dados em clusters, mas também organiza os neurônios de maneira que padrões semelhantes ativem neurônios próximos uns dos outros, criando um mapa topográfico dos padrões de entrada [Fausett, 1994].

De acordo com Ciaburro and Venkateswaran [2017], esse tipo de rede é particularmente útil para visualizar a distribuição e relações entre dados de alta dimensão em um espaço de menor dimensão (geralmente bidimensional), facilitando a interpretação dos agrupamentos e das relações intrínsecas entre diferentes grupos de dados.

Nesta categoria optou-se por utilizar o algoritmo SOM (*Self-Organizing Map*), recomendado para conjuntos de dados multivariados. Este algoritmo foi utilizado com sucesso por Ferlin [2008] em seus estudos, demonstrando bons resultados. O SOM é conhecido por sua capacidade de preservar a topologia dos dados originais, facilitando a visualização e interpretação dos grupos formados. O pseudocódigo do SOM está representado na Figura 4.

Algoritmo 4 Algoritmo de Agrupamento SOM

Require: Dataset (D), Nós ($nodes$), amostra de entrada ($InputSample$), taxa de aprendizado ($LearningRate$), raio de vizinhança ($NeighborhoodRadius$)
Ensure: Labels dos clusters para todos os pontos do conjunto de dados
1: Inicializar os vectores de peso dos *nodes* do SOM com pequenos valores aleatórios.
2: **for** Cada iteração ou até à convergência **do**
3: Selecionar uma amostra aleatória dos dados de entrada.
4: Achar o Best Matching Unit (BMU) - o *node* do SOM com o vetor de peso mais próximo da $InputSample$
5: Ajustar os pesos da BMU e de seus vizinhos para torná-los mais parecidos com a amostra de entrada.
6: Os ajustes diminuem com a distância da BMU e ao longo do tempo.
7: **end for**
8:
9: **function** AdjustWeights(*node*, $InputSample$, $LearningRate$, $NeighborhoodRadius$)
10: NewWeight = OldWeight + $LearningRate$ * ($InputSample$ - OldWeight)
11: Aplicar este ajustamento de peso à BMU e aos seus vizinhos dentro do $NeighborhoodRadius$
12: **end function**

II.4.2 Imputação

A tarefa de imputação desempenha um papel fundamental no processo de preparação e análise de dados, especialmente quando se trata de valores ausentes em conjuntos de dados. Portanto, a imputação de dados ausentes é um passo crítico para garantir a integridade e a utilidade dos dados analisados.

Para dados numéricos, foco desse estudo, diversas técnicas de imputação podem ser aplicadas, com o objetivo de estimar os valores ausentes de maneira a minimizar o impacto na distribuição geral dos dados e preservar, tanto quanto possível, as relações originais entre os atributos. Entre essas técnicas se destacam média ou mediana e algoritmo dos k-vizinhos mais próximos (kNN).

As imputações por média é a técnica mais simples e amplamente utilizada para imputação de dados numéricos [Cartwright et al., 2003]. Consiste em substituir os valores ausentes pela média dos valores disponíveis do mesmo atributo. Embora seja fácil de aplicar, essa abordagem pode não ser ideal em casos de distribuições assimétricas; nesses casos, pode-se utilizar variações como a imputação pela mediana ou pela moda, dependendo da natureza dos dados. Outra desvantagem é a redução da variância, pois todos os valores imputados serão idênticos à média [Cartwright et al., 2003].

Outra técnica de imputação comumente usada é com base em k vizinhos mais próximos (*k-Nearest Neighbors* - kNN). A principal ideia por trás do k-NN é determinar a classificação ou o valor de um dado objeto com base na proximidade desse objeto a outros exemplos previamente classificados ou avaliados no conjunto de dados [Cartwright et al., 2003].

O algoritmo k-NN opera identificando os k casos mais próximos ao objeto que necessita de

imputação, sendo essa proximidade avaliada por similaridade ou distância euclidiana [Soares, 2007]. Para os atributos faltantes, o valor é inferido utilizando as informações dos casos mais semelhantes identificados. Dependendo do número de casos utilizados para a imputação, diferentes estratégias podem ser adotadas, como a média aritmética simples ou um método de ponderação pela distância inversa, como descrito por [Cartwright et al., 2003].

Uma das principais vantagens desta técnica reside na capacidade de permitir que os casos mais similares exerçam maior influência na imputação. Além disso, a flexibilidade em aceitar diversas métricas de distância possibilita a combinação de atributos de características contínuas e categóricas.

Tipos de Distâncias

O agrupamento, uma técnica de aprendizado de máquina não supervisionado, é empregado para organizar dados não etiquetados em grupos, nos quais os pontos de dados compartilham semelhanças entre si e se diferenciam daqueles em outros grupos [Ahmad and Khan, 2019]. As estratégias para medir a proximidade são consideravelmente influenciadas pelas características dos atributos presentes nos dados. De acordo com Ahmad and Khan [2019], os conjuntos de dados do mundo real são categorizados em três tipos principais, com base na natureza de seus valores ao longo das características: numéricos, categóricos e mistos.

No caso de atributos numéricos, a determinação da proximidade entre objetos de dados é mais direta, visto que várias operações numéricas significativas podem ser diretamente empregadas nesses dados para o cálculo da distância [Kar et al., 2023]. Ao longo do tempo, várias medidas de proximidade foram sugeridas para dados numéricos, incluindo a distância Euclidiana, de Manhattan e de *Mahalanobis*, entre outras.

Por outro lado, a computação da (dis)similaridade entre objetos de dados se torna consideravelmente mais complexa quando se trata de atributos categóricos. A principal dificuldade com os conjuntos de dados categóricos reside no fato de seus atributos serem compostos por valores nominais não ordenados. Por essa razão, as medidas de proximidade convencionais, definidas para dados numéricos, não são aplicáveis a eles [Kar et al., 2023]. Para dados categóricos, métricas específicas, como a distância de *Hamming* ou o coeficiente de *Gower*, são mais apropriadas. A Tabela II.1 destaca as principais características das métricas de (dis)similaridades descritas nesta seção, onde: x é um vetor multivariado composto por $x_1, x_2, x_3, ..., x_n$, y é um vetor multivariado composto por $y_1, y_2, y_3, ..., y_n$, $w_{x,y,k}$ são os pesos não-negativos atribuídos entre os x e y em relação à sua k-ésima variável, $s_{x,j,k}$ é semelhança entre os x e y em relação à sua k-ésima variável e S é a matriz covariância da distância entre x e y.

A distância Euclidiana, que calcula a distância linear entre dois pontos no espaço Euclidiano, é intuitivamente compreendida como uma métrica natural do espaço físico [Deza and Deza, 2009].

Tabela II.1: Características das métricas de (dis)similaridades. Fonte: Adaptação de Deza and Deza [2009] e Coombes et al. [2021].

Distância	Tipo de dados	Expressão matemática
Euclidiana	contínuo	$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Manhattan	contínuo	$d(x,y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Mahalanobis	contínuo	$d(x,y) = \sqrt{(x-y)^\top S^{-1}(x-y)}$
Hamming	binário	$d(x,y) = \mathbf{x} - \mathbf{y}$
Gower	nominal, binário e contínuo	$d(x,y) = (\sum_{k=1}^{n} w_{x,y,k} \times s_{x,j,k}) / \sum_{k=1}^{n} w_{x,y,k}$

A distância de Manhattan, também conhecida como métrica do táxi, baseia-se em uma abordagem retilínea, calculada pela soma do valor absoluto das diferenças entre os pontos [Deza and Deza, 2009]. A distância de Mahalanobis, por sua vez, leva em consideração as correlações entre os atributos dos dados, diferenciando-se da distância Euclidiana por ser invariante à escala das medições [Deza and Deza, 2009].

No contexto de dados binários, que são um subtipo de dados categóricos, a distância de Hamming mostra-se particularmente adequada, uma vez que compara as diferenças nas posições dos bits entre vetores binários [Deza and Deza, 2009]. O coeficiente de similaridade de Gower, por sua vez, é recomendado para dados de tipo misto [Coombes et al., 2021], proporcionando uma abordagem holística para a avaliação da (dis)similaridade em conjuntos de dados que integram tanto atributos numéricos quanto categóricos.

Algoritmo k-NN

O algoritmo k-vizinhos mais próximos (*k-Nearest Neighbors* ou k-NN) é uma estratégia fundamental no campo de aprendizado de máquina, que identifica os k elementos mais próximos de um objeto não classificado dentro de um conjunto de treinamento. Esta técnica, detalhada em obras como Aha et al. [1991] e Mitchell [1997], emprega uma métrica de distância — comumente a distância Euclidiana — como indicador de similaridade para agrupar objetos com características semelhantes. É uma técnica frequentemente aplicada em tarefas de classificação e imputação.

Como ferramenta de classificação, o k-NN opera identificando, dentre um conjunto de n tuplas, as k tuplas mais próximas a uma tupla i de classe desconhecida. A partir desse subconjunto, a classe mais frequente (moda) é atribuída à tupla i. A similaridade é geralmente calculada pela distância Euclidiana (seção II.4.2) .

O desafio dessa técnica é a escolha de k, pois variações nesse valor podem alterar o resultado da classificação, como mostra a Figura II.8. Por exemplo, se quiséssemos classificar a observação verde com base em $k = 4$, a observação seria classificada como azul. Se estendermos isso para $k = 9$, a observação passaria a ser classificada como vermelho.

Além da questão da escolha do valor de k, o algoritmo k-vizinhos mais próximos (k-NN) pode

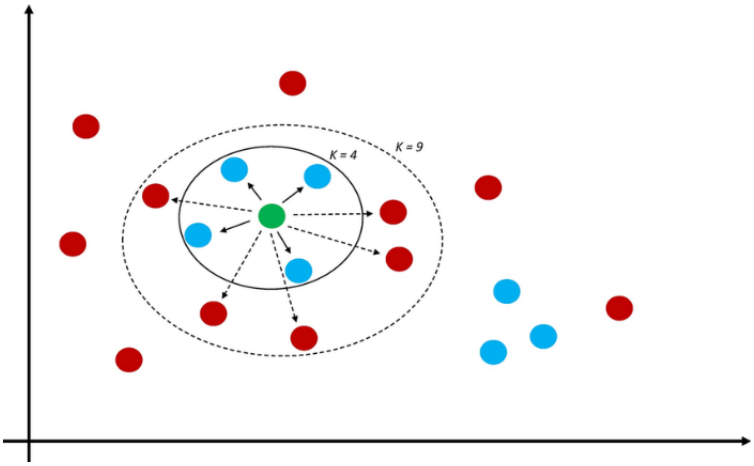


Figura II.8: k-vizinhos mais próximos (Fonte: Musolf et al. [2022])

enfrentar dificuldades na classificação quando os atributos de classes distintas apresentam valores muito semelhantes. No entanto, em cenários onde existe uma clara distinção entre os atributos das diferentes classes, o k-NN tende a ser eficaz.

Uma variante do método permite definir um limiar t (onde $0 < t \leq k$) que estabelece um critério adicional para classificação: uma nova tupla é classificada apenas se, dentre os k vizinhos mais próximos, pelo menos t tuplas pertencem a uma mesma classe C_i. Se esse critério não for atendido, a tupla em questão não é classificada.

Quanto ao uso do k-NN na imputação de dados, este pode ser adaptado como um regressor para atributos contínuos. A técnica envolve o uso da média dos k vizinhos mais próximos ao invés da moda para determinar o valor a ser imputado. O algoritmo k-NN na imputação de uma tupla t_i é apresentado a seguir no pseudocódigo 4

Algoritmo 5 : k-Nearest Neighbors para imputação de valores faltantes

Require: Um conjunto de dados (D) com o atributo t_k preenchido, uma tupla t_i com o atributo em falta t_{ik} e o número k de vizinhos da tupla t_i
Ensure: Tupla t_i com o atributo t_{ik} preenchido
1: Calcular a distância da tupla t_i em todas as outras tuplas t_j no conjunto de dados, $j \neq i$.
2: Ordenar as tuplas no conjunto de dados por ordem crescente da distâncias calculadas.
3: Selecionar os k primeiros tuplas da lista ordenada
4: Atribuir à tupla t_i a média dos atributos t_{jk} das k tuplas seleccionadas.

Soares [2007] destacou em seu trabalho algumas vantagens e desvantagens do algoritmo k-NN que pode ser vistas abaixo:

- Vantagens:
  - É robusto, inclusive para dados com ruídos;
  - Funciona bem mesmo com grandes volumes de dados;

  - Pode regredir atributos tanto categóricos quanto numéricos;
  - Não requer a criação de modelos explícitos para cada atributo com dados ausentes

- Desvantagens:
  - Elevado custo computacional, pois exige o cálculo da distância de cada tupla em relação a todas as outras;
  - A escolha do valor de k pode ser desafiadora e influencia significativamente os resultados;
  - A performance do algoritmo depende fortemente da métrica de distância utilizada;
  - A seleção dos atributos usados no cálculo da distância entre duas tuplas é crucial para o sucesso da aplicação do algoritmo

Jonsson and Wohlin [2004] discutem a superioridade do algoritmo dos k-vizinhos mais próximos (k-NN) sobre métodos tradicionais de imputação, como a média, destacando que a imputação via k-NN é impactada apenas por um subconjunto específico de registros—aqueles mais semelhantes à tupla com valor ausente. Este método é particularmente eficaz pois não leva em consideração todos os registros da base de dados, mas sim aqueles que são estatisticamente relevantes para a imputação em questão.

II.4.3 Métrica

Nesta seção, descreveremos as métricas utilizadas para avaliar os resultados obtidos, focando não apenas nos erros gerados pela imputação, mas também no impacto global que a imputação tem na correlação entre os atributos.

Medida de Erros da Imputação

As principais escolhas como medidas de erro na literatura são o MAE (*Mean Absolut Error* ou Erro Absoluto Médio) e o MAPE (*Mean Absolute Percentage Error*.

O erro absoluto médio é calculado por:

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}|o_i - p_i| \tag{II.1}$$

enquanto que o erro percentual médio absoluto é calculado da seguinte forma:

$$MAPE = \frac{1}{n}\Sigma_{i=1}^{n}\left|\frac{o_i - p_i}{o_i}\right| \times 100\% \tag{II.2}$$

onde, o é o valor observado e p é o valor predito (no nosso caso, valor imputado). Essas medidas indicam a magnitude do erro gerado por um modelo de forma absoluta ou por percentual, fornecendo

uma avaliação clara da diferença entre os valores previstos e os valores observados (Figura II.9), expressando o tamanho típico do erro em unidades comparáveis aos valores dos dados [Willmott, 1981].

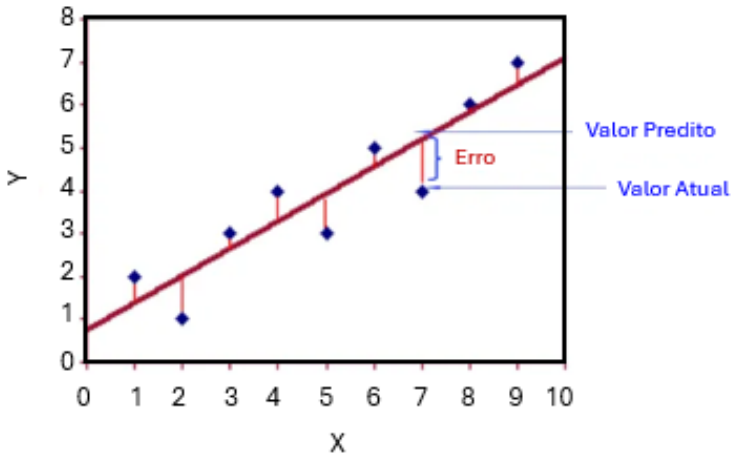


Figura II.9: Visualização do erro de imputação calculado entre a diferença do valor observado (o) e o valor predito (p) (Fonte: elaboração própria)

Soares [2007] modificou uma medida de similaridade originalmente desenvolvida por Magnani [2004], adaptando-a para funcionar como uma medida de erro. Esta medida, projetada para avaliar a semelhança entre registros em um conjunto de dados, é aplicada individualmente a cada atributo dos registros. A função especifica o grau de similaridade entre registros, quantificando o quão próximos ou equivalentes eles são com base em seus atributos correspondentes (Equação II.3).

$$sim(a_k, b_k) = \frac{|a_k - b_k|}{|max_x(k) - min_x(k)|} \tag{II.3}$$

onde, a e b são duas tuplas de uma tabela, o k indica o atributo k, $max_x(k)$ e $min_x(k)$ são, respectivamente, o maior e o menor valores apresentados na tabela para a coluna k para todas as tuplas x da tabela.

Se tomarmos a_k como o valor original e b_k como o valor imputado, essa métrica pode ser vista como uma medida de erro que quantifica a distância entre o valor imputado e o original. Essa medida, denominada "erro de similaridade" (Equação II.4), foi a escolhida para avaliar a eficácia da imputação. Ela ajuda a entender o quão próximo o valor imputado está do valor verdadeiro, fornecendo uma base quantitativa para avaliar o desempenho do método de imputação utilizado.

$$E_{sim} = \frac{|x - x'|}{|max_x - min_x|} \tag{II.4}$$

onde, x é o valor original e x' é o valor imputado.

Para calcular o erro de similaridade médio de um processo de imputação(Equação II.5), utiliza-

mos a média aritmética dos erros de similaridade individuais de cada atributo ou de cada grupo de atributos imputados. A fórmula para o erro médio de similaridade é expressa da seguinte maneira:

$$\overline{E_{sim}} = \frac{1}{n}\Sigma_{i=1}^{n} E_{sim}(i) \tag{II.5}$$

Essa métrica é aplicada para avaliar o processo de imputação em diferentes contextos, seja para um único atributo, para um conjunto de atributos dentro de um mesmo agrupamento, ou para o conjunto total de dados. Dessa forma, é possível quantificar e comparar a eficiência da imputação em diferentes níveis de granularidade da análise.

Medida de Preservação da Correlação de Atributos

A natureza recursiva do processo de imputação em cascata pode intensificar as relações intrínsecas entre os atributos de um conjunto de dados, pois os valores imputados são derivados dos dados existentes e subsequentemente tratados como se fossem originais para as rodadas seguintes de imputação. Esse mecanismo de realimentação é análogo ao conceito financeiro de "juros compostos", onde os ganhos de uma etapa são reinvestidos para ampliar os resultados futuros.

O coeficiente de correlação entre dois atributos X e Y é dado por:

$$\rho X, Y = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{II.6}$$

onde *cov(X, Y)* é a matriz de covariância entre os atributos X e Y, σ_X é o desvio padrão do atributo X e σ_Y do atributo Y. O coeficiente de correlação é uma medida estatística que varia entre -1 e 1 [Bussab and Moretin, 2017], refletindo a força e a direção da relação linear entre dois atributos. Um coeficiente de -1 indica uma correlação negativa perfeita (onde os valores de um atributo aumentam enquanto os do outro diminuem), e um coeficiente igual a 1 mostra uma correlação positiva perfeita (onde os valores de um atributo aumentam em uníssono com os do outro). Esta métrica é essencial para entender como os atributos de um conjunto de dados interagem entre si, especialmente em análises onde a interdependência dos dados é crucial para a modelagem.

A matriz de correlação em um conjunto de dados é uma ferramenta que apresenta uma estrutura quadrada $n \times n$, sendo n o número de atributos do conjunto. Cada elemento $M(i, j)$ desta matriz representa a correlação entre o atributo k_i e o atributo k_j, ilustrando a relação linear entre eles.

$$\begin{bmatrix} \rho(k_1, k_1) & \rho(k_1, k_2) & \dots & \rho(k_1, k_n) \\ \rho(k_2, k_1) & \rho(k_2, k_2) & \dots & \rho(k_2, k_n) \\ \vdots & \ddots & \vdots & \\ \rho(k_n, k_1) & \rho(k_n, k_2) & \dots & \rho(k_n, k_n) \end{bmatrix}$$

A correlação total de um atributo k pode ser calculada como a média das correlações desse atributo com todos os outros atributos no conjunto. Por exemplo, a correlação total do atributo k_1 pode ser visualizada na Equação II.7 como:

$$OC(k_1) = \frac{\rho(k_1, k_2) + \rho(k_1, k_3) + ... + \rho(k_1, k_n)}{n - 1} \tag{II.7}$$

Um dos objetivos principais deste estudo é analisar como a imputação afeta a correlação entre os atributos. Para isso, a diferença entre a correlação total de um atributo na base imputada e na base original é calculada. Essa diferença, chamada de ACB (*Attribute Correlation Bias* ou Bias da Correlação do Atributo), indica o desvio introduzido pela imputação na correlação desse atributo (Equação II.8).

$$ACB(k, k') = OC(k) - OC(k') \tag{II.8}$$

, onde k é o valor original e k' é o valor imputado.

Como uma forma de avaliar o impacto global da imputação na correlação entre os atributos de um conjunto de dados, o somatório dos ACBs de todos os atributos é calculado. Este valor agregado é denominado Bias da Correlação (*Correlation Bias* - CB) e serve como uma métrica global para quantificar as alterações na estrutura de correlação devido ao processo de imputação (Equação II.9).

$$CB = \Sigma_{i=1}^{n} ACB(k_i, k_i') \tag{II.9}$$

II.4.4 Teste de Hipóteses

O teste de hipóteses é uma técnica estatística usada para tomar decisões ou inferências sobre parâmetros populacionais com base em dados amostrais. O objetivo do teste é, portanto, fornecer uma metodologia que possibilite verificar se os dados amostrais oferecem evidências que sustentem ou refutem uma hipótese estatística formulada [Bussab and Moretin, 2017].

Existem dois tipos principais de testes de hipóteses: paramétricos e não paramétricos. Os testes paramétricos são baseados em suposições sobre a distribuição dos dados, ou seja, eles assumem que a amostra segue uma distribuição específica, geralmente a distribuição normal. Os testes não paramétricos, por outro lado, não fazem suposições rigorosas sobre a distribuição dos dados.

Denomina-se hipótese nula (H_0)uma proposição formulada com o objetivo de ser testada. A hipótese alternativa (H_1) é a proposição considerada válida caso as informações obtidas da amostra resultem na rejeição da hipótese nula.

Como a decisão de aceitar ou rejeitar H_0 é baseada unicamente nas informações provenientes de uma amostra da população, existe a possibilidade de cometer um dos seguintes tipos de erro:

- Erro do tipo I ou erro de primeira espécie: ocorre quando se rejeita uma hipótese nula verdadeira.
- Erro do tipo II ou erro de segunda espécie: ocorre quando se aceita uma hipótese nula falsa.

Dado que esses erros são inevitáveis, a boa prática em testes estatísticos é minimizar a probabilidade de ocorrência de ambos. Entretanto, como o erro do tipo I é geralmente considerado mais grave, o foco principal nos testes de hipóteses está em controlar e reduzir a probabilidade de cometer esse tipo de erro.

Teste de Wilcoxon

< REVISAR >

O teste de *Wilcoxon* é um teste estatístico não paramétrico que serve para comparar duas amostras dependentes ou emparelhadas. Ele é particularmente útil quando se têm dados pareados e deseja-se verificar se há diferenças significativas entre as distribuições de dois conjuntos de dados relacionados. Diferentemente de outros testes, como o teste t de *Student*, o teste de *Wilcoxon* não exige que as amostras sigam uma distribuição normal, o que o torna adequado para situações em que essa suposição não é válida.

Para realizar o teste, considere que X_i representa os dados originais e Y_i os valores imputados. Inicialmente, calcula-se a diferença para cada par de escores, ou seja, $d_i = X_i - Y_i$. Após o cálculo das diferenças, atribui-se postos às diferenças absolutas dos escores, $|d_i|$. A menor diferença em valor absoluto recebe o posto 1, a segunda menor recebe o posto 2, e assim por diante. Em seguida, o sinal das diferenças originais (d_i) é adicionado aos postos.

Sob a hipótese nula (H_0), espera-se que as somas dos postos positivos ("+") e negativos ("–") sejam aproximadamente iguais, indicando que não há diferença entre as distribuições dos dados. Caso as somas não sejam aproximadamente iguais, pode-se concluir que as distribuições dos dados são diferentes, levando à rejeição de H_0(ou seja, aceitamos H_1). Para simplificar, nossas hipóteses foram:

H_0: as distribuições dos dados originais e imputados são iguais.
H_1: as distribuições dos dados originais e imputados são diferentes.

Empates podem ocorrer de duas formas: quando a diferença entre os escores X_i e Y_i for zero ($|d_i| = 0$), o par é excluído da análise. Quando existem diferenças absolutas $|d_i|$ iguais, atribui-se a essas diferenças a média dos postos que elas receberiam caso não fossem empatadas. Em seguida, o sinal das diferenças originais é adicionado aos postos. Por exemplo, se as menores diferenças entre os escores forem $d_1 = 1$, $d_2 = -1$ e $d_3 = -1$, as diferenças absolutas seriam $|d_1| = 1$, $|d_2| = 1$ e $|d_3|$

= 1. Nesse caso, o posto 2 seria atribuído a cada uma das três diferenças, uma vez que a média de 1, 2 e 3 é 2. Assim, os postos com os sinais das diferenças seriam: 2, -2 e -2. A próxima diferença na ordem receberia o posto 4, e assim sucessivamente.

No caso de amostras grandes ($n > 20$), pode-se utilizar a aproximação da distribuição dos valores críticos de V pela distribuição Normal. Nesse caso, a soma dos postos V, tem distribuição aproximadamente normal com média (Equação II.10):

$$\mu_0(V) = \frac{n(n+1)}{4} \tag{II.10}$$

e desvio padrão (Equação II.11)

$$\sigma_0(V) = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{II.11}$$

A estatística de teste é dada por (Equação II.12) segue uma distribuição normal padrão.

$$Z_v = \frac{V - \mu_0}{\sigma_0} \tag{II.12}$$

Como V é uma variável aleatória discreta, deve-se usar a correção de continuidade. Portanto, a estatística do teste é dada por (Equação II.13):

$$Z_v = \frac{(v \pm 0,5) - \mu_0}{\sigma_0} = \frac{(v \pm 0,5) - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{II.13}$$

A determinação do *p-value* no teste de *Wilcoxon* é realizada com base na distribuição da estatística do mesmo. O *p-value* indica a probabilidade de observar uma estatística de teste tão extrema quanto a calculada, assumindo que a hipótese nula (H_0) seja verdadeira. Caso o *p-value* seja inferior ao nível de significância estabelecido (geralmente 0,05), rejeita-se a hipótese nula, sugerindo que há uma diferença estatisticamente significativa entre os dois conjuntos de dados emparelhados.

Capítulo III Trabalhos relacionados

Este capítulo abrange a revisão dos estudos que são fundamentais para o desenvolvimento do projeto de pesquisa atual, enfatizando o estudo-chave que formam a espinha dorsal desta investigação: Soares [2007] e Ferlin [2008]. A revisão da literatura sobre a temática de imputação em cascata foi realizada seguindo as normas estabelecidas pelo PRISMA [Page et al., 2021].

III.1 Seleção de trabalhos

Para iniciar a busca por literatura relevante, foi elaborada uma estratégia de pesquisa em inglês, composta por palavras-chave e seus sinônimos associados ao tema do estudo. Os termos principais utilizados foram *"missing data"* e *"imputation"*. Além desses, foram incorporados à estratégia de busca outros termos que alinhassem mais estreitamente com o foco específico da pesquisa, resultando nas seguintes expressões de busca: *"cascade imputation" OR "cascade missing data" OR "recursive imputation" OR "recursive missing data" OR "iterative imputation" OR "iterative missing data"*. A pesquisa incluiu trabalhos em inglês, sem restrições quanto ao período de publicação.

Para o acesso aos estudos, optou-se pela utilização da base de dados Scopus Elsevier, reconhecida pela sua ampla cobertura de publicações científicas.

III.1.1 Critérios de inclusão

A aplicação de critérios de inclusão são fundamentais para assegurar a relevância e qualidade da revisão da literatura. Para este projeto, foram estabelecidos dois critérios de inclusão específicos:

- **Critério 1:** Consideramos artigos de revista e conferências;
- **Critério 2:** Sem limite de tempo.

III.1.2 Critérios de exclusão

Os critérios de exclusão são essenciais para delimitar o escopo da pesquisa ao foco desejado, descartando estudos que não atendem aos objetivos propostos. Para este projeto, foram estabelecidos cinco critérios de exclusão específicos:

- **Critério 1:** Exclusão de trabalhos que discutem temas biológicos ou genéticos;

- **Critério 2:** Exclusão de estudos que abordam a imputação em dados de imagem, como bytes ausentes de uma imagem;
- **Critério 3:** Exclusão de pesquisas que tratam de dados de áudio ou som;
- **Critério 4:** Exclusão de trabalhos focados na imputação de textos; e
- **Critério 5:** Exclusão de trabalhos que não estejam em sua versão final publicada.

III.1.3 String de Busca

Com base nos critérios de inclusão (III.1.1) e exclusão (III.1.2), o descritor completo para a realização da busca foi definido como: (TITLE-ABS-KEY(“cascade imputation” OR “cascade missing data”) OR TITLE-ABS-KEY(“recursive imputation” OR “recursive missing data”) OR TITLE-ABS-KEY(“iterative imputation” OR “iterative missing data”) AND NOT TITLE-ABS-KEY (“microarray” OR “RNA” OR “gene” OR “gene expression profiling” OR “gene OR data” OR “genotype” OR “genomics” OR “genetics algorithms” OR “biology” OR “bioinformatics” OR “proteomics"OR “genetics” OR “human” OR “humans” OR “animals”) AND NOT TITLE-ABS-KEY (“image” OR “images”) AND NOT TITLE-ABS-KEY (“signal” OR “signal processing” OR “wireless” OR “audio” OR “sound”) AND NOT TITLE-ABS-KEY (“web” OR “text”)) AND (LIMIT-TO (PUBSTAGE,“final”))

III.1.4 Resultado da Busca

A pesquisa de trabalhos relacionados foi conduzida na base Scopus, o principal repositório de artigos da área de Ciências da Computação, utilizando a estratégia de busca elaborada com as palavras-chave selecionadas, a pesquisa foi realizada em 17/06/2024. A busca realizada na base Scopus retornou 43 trabalhos, dos quais 11 não foram acessíveis devido a erros nos links, redirecionamentos repetidos e exigências de assinatura para acesso aos textos completos. Além disso, 5 dos artigos estavam fora do escopo ou eram teóricos; 12 artigos aplicaram métodos específicos para imputação de dados, possibilitando ou não posterior modelagem (como regressão ou classificação); 4 artigos se limitaram a comparar resultados entre diferentes técnicas de imputação; 10 artigos utilizaram métodos iterativos sem reincorporar atributos imputados anteriormente para auxiliar na imputação de atributos subsequentes. Por fim, um artigo, desenvolvido por Montiel et al. [2018], indica que a imputação em cascata é ainda pouco explorada na literatura acadêmica, utilizando este termo ou seus equivalentes, ressaltando a relevância e as oportunidades para futuros desenvolvimentos nesta área de pesquisa.

O artigo intitulado “*Scalable Model-Based Cascaded Imputation of Missing Data*” [Montiel et al., 2018], introduz o método de *Cascade Imputation* (CIM), uma técnica incremental baseada em

aprendizado de máquina supervisionado. Este método é detalhado através de uma sequência de etapas metodológicas: (1) *Ordenação do Conjunto de Dados*: inicialmente, o conjunto de dados é organizado em ordem crescente, com base na quantidade de dados ausentes em cada atributo, facilitando a aplicação sequencial dos passos subsequentes; (2) *Separação e Aplicação de Modelos*: o conjunto de dados é dividido em subconjuntos de treino e teste. Dependendo do tipo de variável-alvo (numérica ou categórica), um modelo de regressão ou de classificação é aplicado, respectivamente. O conjunto de treino consiste nos atributos que possuem dados completos, que são então utilizados para treinar o modelo. Este modelo é posteriormente empregado para prever os valores ausentes no conjunto de teste; (3) *Imputação Iterativa*: na rodada subsequente do processo, o atributo que foi previamente imputado é incluído como uma variável preditora nos modelos de regressão ou classificação. Este ciclo de imputação e reintegração de variáveis continua até que não restem atributos com valores ausentes no conjunto de dados. A Figura III.1 exemplifica como funciona da proposta de Montiel et al. [2018].

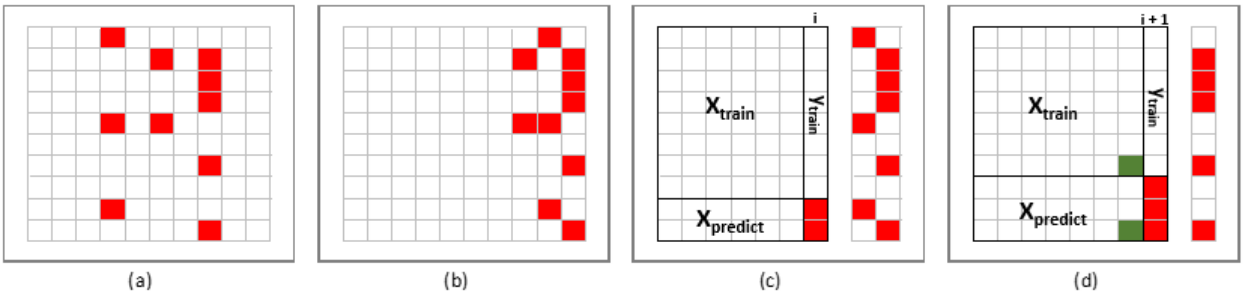


Figura III.1: Representação do Trabalho de Montiel et al. [2018]. (a) Dados originais com valores em falta faltantes em vermelho. (b) Posições atualizadas após ordenação dos atributos por contagem de valores em falta. (c-d) Iterações de imputação, repetidas até o conjunto de dados estar completo. Os valores imputados, em verde, são utilizados nas iterações seguintes do algoritmo. (Fonte: Montiel et al. [2018])

Os trabalhos de Soares [2007] e Ferlin [2008] foram a base desse projeto. A abordagem de imputação composta, desenvolvida por Soares [2007], apresenta a ideia da aplicação de técnicas de imputação anteriores ao processo de imputação propriamente dito. Essas técnicas de aprendizado de máquina, como seleção, agrupamento ou classificação, buscam aprimorar a precisão dos dados imputados. Esta metodologia é fundamentada no conceito de planejamento estratégico e de imputação, facilitando a execução organizada e metódica da imputação de dados. O trabalho de Soares [2007] sugeriu, como direção para futuras pesquisas, o desenvolvimento da imputação em cascata; esta recomendação influenciou indiretamente o presente estudo.

Essa sequência de atividades preparatórias é conhecida como estratégia e plano de imputação, onde cada tarefa específica é associada a um algoritmo apropriado. Como exemplo, uma estratégia típica poderia incluir etapas como: (i) seleção de atributos significativos; (ii) agrupamento de casos; e (iii) imputação dos valores ausentes. Soares [2007] experimentalmente observou que a implementa-

ção de uma etapa de agrupamento antes da imputação traz melhorias significativas em comparação com métodos que realizam a imputação diretamente, destacando a necessidade de explorar essa técnica com algoritmos inovadores.

Nos estudos conduzidos por Soares [2007], foram utilizados dados de três diferentes conjuntos do *UCI Machine Learning Repository*. Em cada conjunto, foram produzidas artificialmente ausências variando de 10% a 50% segundo o mecanismo de ausência MCAR (*Missing Completely At Random*). A estratégia empregada envolveu o arranjo em workflow das seguintes tarefas: (i) seleção de atributos, usando o algoritmo PCA; (ii) agrupamento, utilizando o algoritmo K-means; e (iii) imputação dos valores ausentes utilizando a média dos dados agrupados, ou o algoritmo dos k vizinhos mais próximos (*k-nearest neighbors*), ou a rede neuronal artificial *back propagation*. Esse conjunto de procedimentos definiu uma nova metodologia denominada **imputação composta**, que se caracteriza pela composição em workflow de zero ou mais tarefas prévias à imputação propriamente dita, visando melhorar a qualidade do dado imputado.

Impulsionado pelos resultados positivos de Soares [2007], que demonstrou a eficácia de realizar tarefas de agrupamento antes da imputação - remetendo ao conceito de *imputação hot deck*, Ferlin [2008] desenvolveu um método inovador denominado **imputação em cascata** que descreveremos mais detalhadamente na próxima seção.

III.2 Imputação em Cascata

A imputação em cascata, proposta por Ferlin [2008], surge com o objetivo de melhorar a qualidade do processo de imputação de dados. Nela, os valores imputados em um atributo são reintegrados à base de dados antes da imputação do atributo subsequente, ao invés de serem adicionados somente ao final de todas as iterações, como é comum nesse tipo de procedimento. Esse processo é repetido até que a imputação seja concluída para todo o conjunto de dados.

Este método é estruturado em quatro etapas principais: (i) pré-processamento; (ii) morfologia da ausência; (iii) critério de ordenação de segmentos e atributos; e (iiiv) imputação sequencial com realimentação. O fluxo de trabalho completo está ilustrado na Figura III.2 e será detalhado nas seções subsequentes.

III.2.1 Pré-processamento

A etapa de pré-processamento estabelece a qualidade inicial dos dados a serem analisados. Na imputação em cascata de Ferlin [2008], inicialmente, essa fase envolve a normalização dos valores da base de dados para uniformizar a escala e minimizar desvios causados por diferenças na magnitude dos atributos. Este ajuste é realizado por meio de métodos de normalização, que podem mapear os valores para intervalos padronizados como [0.0, 1.0] ou [-1.0, 1.0]. Importante destacar, conforme

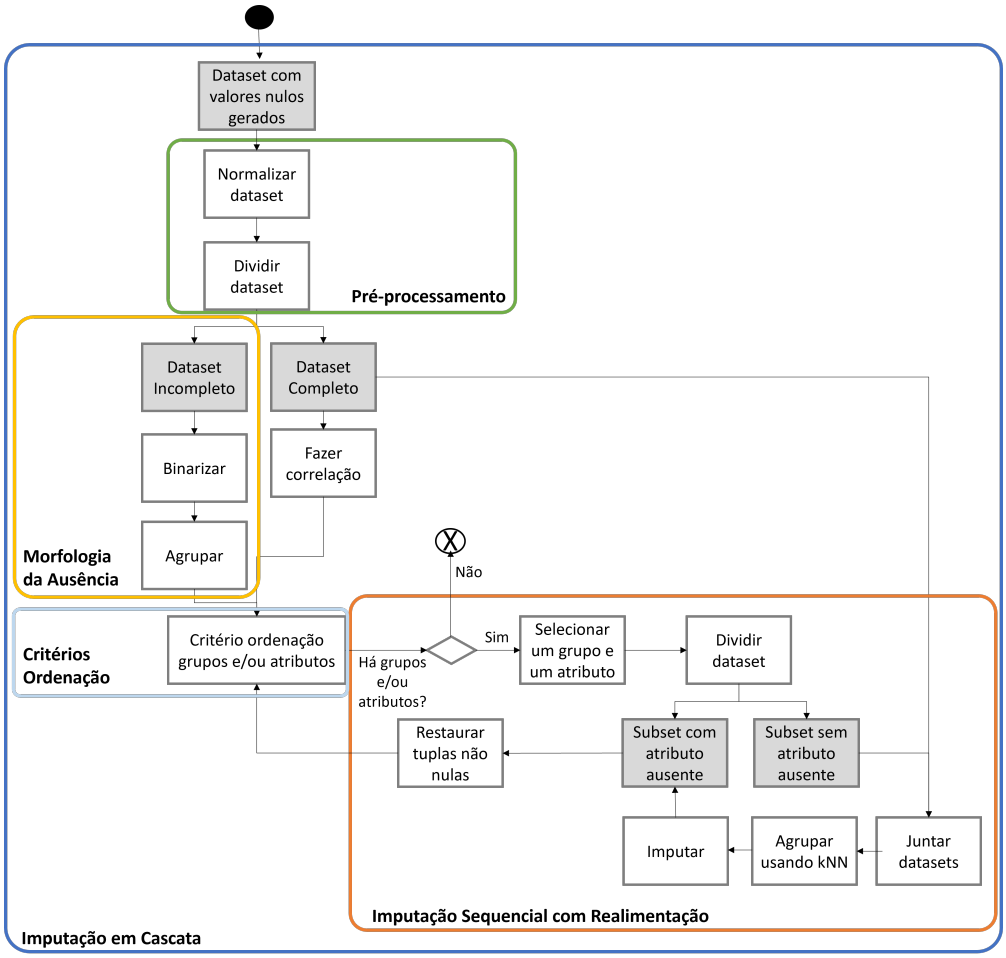


Figura III.2: Fluxo do trabalho proposto por Ferlin [2008]

mencionado pelo autor, que a escolha do método de normalização é flexível e pode ser determinada pelo analista conforme a necessidade da análise. Além disso, ocorre a separação dos dados em duas bases distintas: uma contendo casos completos e outra com casos incompletos.

O subconjunto contendo dados completos foi empregado para calcular a correlação entre os atributos, uma informação que será utilizada na etapa subsequente de definição dos critérios de ordenação. Este mesmo subconjunto constitui a base original inicial para a quarta etapa, que consiste na imputação dos valores faltantes. Já o subconjunto incompleto é a base da segunda etapa que é a da morfologia da ausência.

III.2.2 Morfologia da Ausência

A finalidade desta etapa é organizar os dados com base na distribuição espacial dos valores ausentes no conjunto de dados. A Figura III.3 ilustra o processo desenvolvido nesta etapa.

Este estágio começa com a binarização dos dados incompletos, provenientes da etapa de pré-processamento, no qual os valores ausentes são representados por 1, e os valores presentes por 0 como exemplifica a Figura III.4. Esse processo assegura que o algoritmo de agrupamento selecionado

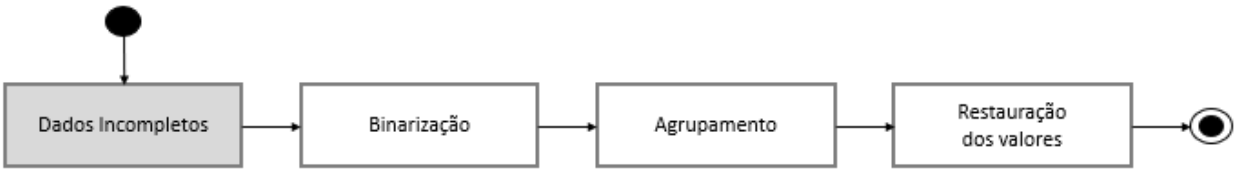


Figura III.3: Representação da etapa da Morfologia da Ausência (Fonte: Adaptação da proposta de Ferlin [2008])

não seja influenciado pelos valores dos atributos, mas sim pela sua configuração espacial.

Base incompleta

Atrib. 1	Atrib. 2	Atrib. 3	Atrib. 4
0,017	-	0,092	0,068
-	0,565	0,141	-
-	-	0,112	0,497
0,113	0,256	-	0,576
0,020	0,081	-	-
-	0,096	0,837	-

(a)

Base binarizada

Atrib. 1	Atrib. 2	Atrib. 3	Atrib. 4
1	0	1	1
0	1	1	0
0	0	1	1
1	1	0	1
1	1	0	0
0	1	1	0

(b)

Figura III.4: Representação da binarização

Como em qualquer tarefa de agrupamento, o objetivo é garantir que as observações dentro de cada grupo sejam homogêneas, enquanto que ao comparar diferentes grupos, estas apresentem heterogeneidade. Nesta tarefa de agrupamento, Ferlin [2008] se restringiu ao uso de redes neurais auto-organizáveis (*Self-Organizing Maps* - SOM). Logo após a segmentação do subconjunto incompleto, procede-se à restauração dos valores, ou seja, os dados são revertidos aos seus estados originais.

III.2.3 Critérios quanto à ordenação dos segmentos e atributos

Ferlin [2008] propôs a hipótese de que a ordenação dos segmentos, criados durante a análise da morfologia da ausência, e dos atributos, têm impacto significativo na qualidade da imputação e, consequentemente, no erro associado. A seguir, são apresentados os critérios de ordenação adotados pela autora.

Ordenação dos segmentos

Uma estratégia para a imputação em cascata, considerando a ordenação dos grupos provenientes da etapa da morfologia da ausência, foi proposto no trabalho de Ferlin [2008]. As abordagens sugeridas para a seleção da ordem em que os segmentos são imputados incluem:

1. ***tupleLessMissing*:** inicia com o grupo que tem o menor número de tuplas ausentes, e progride para o grupo com o maior número;

2. ***tupleMoreMissing*:** contrária à abordagem anterior, inicia com o grupo que tem o maior número de tuplas ausentes e termina com o que tem o menor;

3. ***fieldLessMissing*:** prioriza grupos com menos células ausentes em geral, avançando para aqueles com mais células ausentes;

4. ***fieldMoreMissing*:** na contramão da abordagem anterior, começa com grupos que têm mais células ausentes e vai para os que têm menos;

5. ***fieldPerTupleLessMissing*:** seleciona grupos com base na menor média de células ausentes por tupla, avançando para a maior média;

6. ***fieldPerTupleMoreMissing*:** em oposição à abordagem anterior, inicia com o grupo que tem a maior média de células ausentes por tupla e termina com o menor;

7. ***random*:** a ordem dos grupos é escolhida aleatoriamente;

8. ***noSort*:** utiliza a ordem natural em que os grupos foram gerados durante o processamento.

O objetivo dessa metodologia é equilibrar o volume de dados disponíveis para imputação com a qualidade dos registros imputados. Desta forma, busca-se maximizar a precisão das imputações, adaptando a estratégia ao contexto dos dados e suas características específicas de ausência. Essa técnica permite que as estimativas sejam tão próximas quanto possível dos valores originais, embora isso dependa da abundância e da qualidade dos dados "doadores" disponíveis no conjunto de treinamento.

Ordenação de imputação de um atributo

A imputação multivariada de múltiplos atributos é tratada através de uma sequência de imputações univariadas. Esta abordagem exige a definição de uma ordem específica para a imputação dos atributos, um aspecto que tem um impacto significativo no processo de complementação dos dados. Influenciado por pesquisas anteriores, que frequentemente priorizam os atributos com base na quantidade de dados ausentes, Ferlin [2008] propôs cinco abordagens distintas para a ordenação dos atributos durante a imputação:

1. ***lessCorrelation*:** começa com o atributo que tem a menor correlação (calculada pela soma dos valores absolutos das correlações desse atributo com os outros) e avança até o mais correlacionado;

2. ***moreCorrelation*:** o oposto do primeiro, começando pelo atributo mais correlacionado e progredindo para o menos correlacionado;

3. ***lessMissing*:** inicia com o atributo que possui menos dados ausentes, seguindo para aquele com mais dados ausentes (critério usual);

4. ***moreMissing*:** inverso do critério usual, começando pelo atributo com mais dados ausentes até o com menos dados ausentes;

5. ***noSort*:** adota a ordem original dos atributos conforme definida na estrutura da base de dados.

As estratégias propostas exploram a possibilidade de que características específicas dos atributos, especialmente no que diz respeito aos seus dados faltantes, possam influenciar a qualidade das imputações. Ao integrar essas características na ordem das imputações univariadas, busca-se avaliar se esses detalhes contribuem para uma melhoria significativa na precisão do preenchimento dos valores ausentes.

III.2.4 Imputação Sequencial com Realimentação

Nesta etapa, o objetivo é realizar a imputação de múltiplos atributos ausentes de maneira sequencial, abordando-a como uma série de imputações univariadas. O fluxo de trabalho para esta etapa é detalhado na Figura III.5.

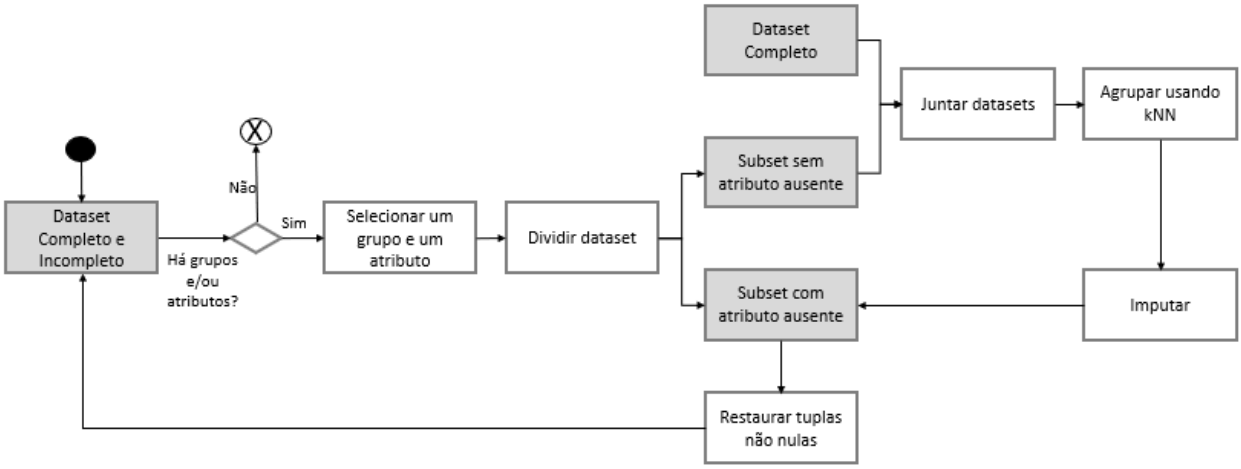


Figura III.5: Fluxo de trabalho da imputação sequencial com realimentação (Fonte: Adaptação da proposta de Ferlin [2008])

Inicialmente os dados incompletos são organizados de acordo com os critérios de ordem dos segmentos (Seção III.2.3) e de atributos (Seção III.2.3). Posteriormente, os dados são subdivididos em duas sub-bases: (i) com atributo ausente e (ii) sem atributo ausente. O conjunto de dados completo e a sub-base sem atributos ausentes são empregados no agrupamento utilizando o método kNN. Durante esta etapa, são identificadas as tuplas mais semelhantes, e os valores são imputados

na sub-base que contém atributos ausentes. Após a imputação, as tuplas não nulas são reintegradas ao conjunto de dados completo.

A imputação é realizada por sub-base (Figura III.5), tratando todas as tuplas incompletas de um atributo específico antes de avançar para o próximo. Este procedimento é repetido até que todos os atributos incompletos nas tuplas sejam restaurados.

No trabalho desenvolvido por Ferlin [2008], é destacado que os parâmetros ajustáveis do algoritmo kNN podem ser definidos pelo analista. As seguintes configurações foram adotadas em seus experimentos: (i) valores de k iguais a 1, 3, 5 e 10; (ii) duas medidas de distância, a Euclidiana e a Mahalanobis; e (iii) para o cálculo do valor estimado, optou-se pela média dos valores dos atributos preenchidos dos vizinhos mais próximos.

III.2.5 Avaliação dos Experimentos

Para avaliar a qualidade da imputação realizada em seus experimentos, Ferlin [2008] empregou várias métricas específicas. Como medidas de erro da imputação, foram escolhidas o erro absoluto médio (MAE - *Mean Absolute Error*), o erro percentual médio absoluto (MAPE - *Mean Absolute Percentage Error*) e a medida de erro de similaridade, conforme descrito na Seção II.4.3. Essas medidas permitiram avaliar a magnitude do erro introduzido pela imputação, tanto de forma absoluta quanto percentual, e possibilitaram comparar a eficiência da imputação em diferentes níveis de granularidade analítica.

Além disso, um dos objetivos do estudo de Ferlin [2008] foi investigar o impacto da imputação na correlação entre os atributos. Para tal análise, foi utilizada a medida de viés de correlação (CB - *Correlation Bias*), que atua como uma métrica global para quantificar as alterações na estrutura de correlação resultantes do processo de imputação e que foi descrita na Seção II.4.3.

III.2.6 Limitações do Trabalho

Algumas limitações e possíveis extensões do trabalho desenvolvido por Ferlin [2008] foram identificadas e listadas, as quais serão consideradas e abordadas no âmbito dessa pesquisa.

- Na fase de agrupamento, considerar a avaliação de outros métodos além das redes neurais auto-organizáveis, explorando abordagens hierárquicas ou até mesmo determinísticas, baseadas no padrão de ausência dos dados.

- Para a análise da eficácia da abordagem, é proposto expandir os mecanismos de ausência de dados utilizados, incluindo não apenas o MCAR (*Missing Completely At Random*), mas também o NMAR (*Not Missing At Random*) e o MAR (*Missing At Random*). Adicionalmente, recomenda-se avaliar o desempenho da metodologia em grandes conjuntos de dados.

- Proceder a formalização do processo de imputação em cascata, com o objetivo de tornar o método mais abstrato e aplicável em diversos contextos de pesquisa.

III.2.7 Comparativo dos Trabalhos de Ferlin e Montiel

O estudo de Ferlin [2008] se diferencia da abordagem de Montiel et al. [2018] por implementar a imputação *hot-deck* - na qual a imputação considera apenas as linhas de dados com maior similaridade no processo de imputação, ao invés da imputação clássica, que leva em conta todo o conjunto de dados nesse processo - com algoritmos de agrupamento de quatro categorias distintas (particionamento, hierárquico, modelo e densidade), além de realizar o processo de imputação com um algoritmo (k-NN) que não demanda a divisão do *dataset* em conjunto de treino e teste.

Capítulo IV Metodologia

Este capítulo discute a metodologia empregada neste trabalho, incluindo a seleção das bases de dados, os algoritmos escolhidos para a execução dos testes, e a configuração do ambiente computacional. A escolha cuidadosa desses elementos é essencial para assegurar a validade e a eficácia dos resultados obtidos nos testes. A discussão detalhada de cada componente metodológico proporcionará clareza sobre as técnicas e as ferramentas aplicadas, estabelecendo uma base sólida para a interpretação dos resultados e a comparação dos diferentes métodos de imputação testados neste estudo.

IV.1 Descrição da Abordagem Proposta do Trabalho

Este trabalho é uma extensão da pesquisa realizada por Ferlin [2008], porém incorpora abordagens distintas que serão detalhadas nesta seção. O estudo em questão compara três métodos: (i) imputação univariada, (ii) imputação em cascata e, (iii) imputação em cascata simplificado (Figura IV.1). Embora a geração inicial de valores faltantes através dos mecanismos de ausência e a subsequente separação da base de dados em conjuntos completo e incompleto sejam comuns aos três métodos, como pode ser observado na Figura IV.1, o método de imputação em cascata difere por incluir uma etapa adicional de binarização dos dados incompletos antes da imputação propriamente dita.

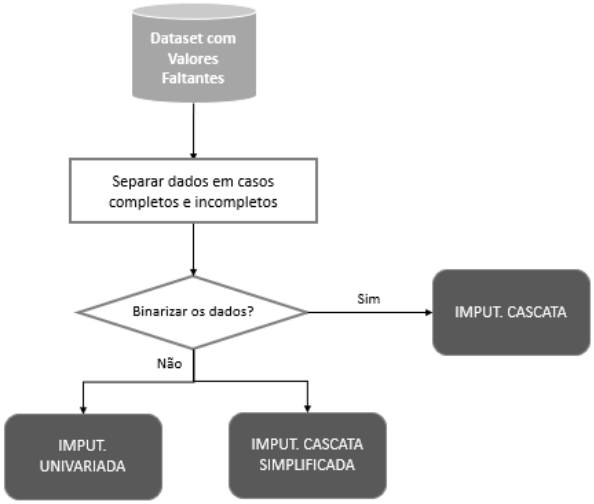


Figura IV.1: Visão simplificada da abordagem da proposta do estudo

A metodologia denominada “Imputação em Cascata” desse trabalho adota o mesmo princípio usado por Ferlin [2008] (Seção III.2), investigando se a segmentação prévia dos dados com base no padrão de ausência — ou seja, na distribuição espacial dos valores faltantes nas tuplas — pode melhorar a eficácia da imputação. Este procedimento explora a hipótese de que entender e agrupar os dados conforme a sua morfologia de ausência antes da imputação pode aprimorar significativamente os resultados da imputação subsequente.

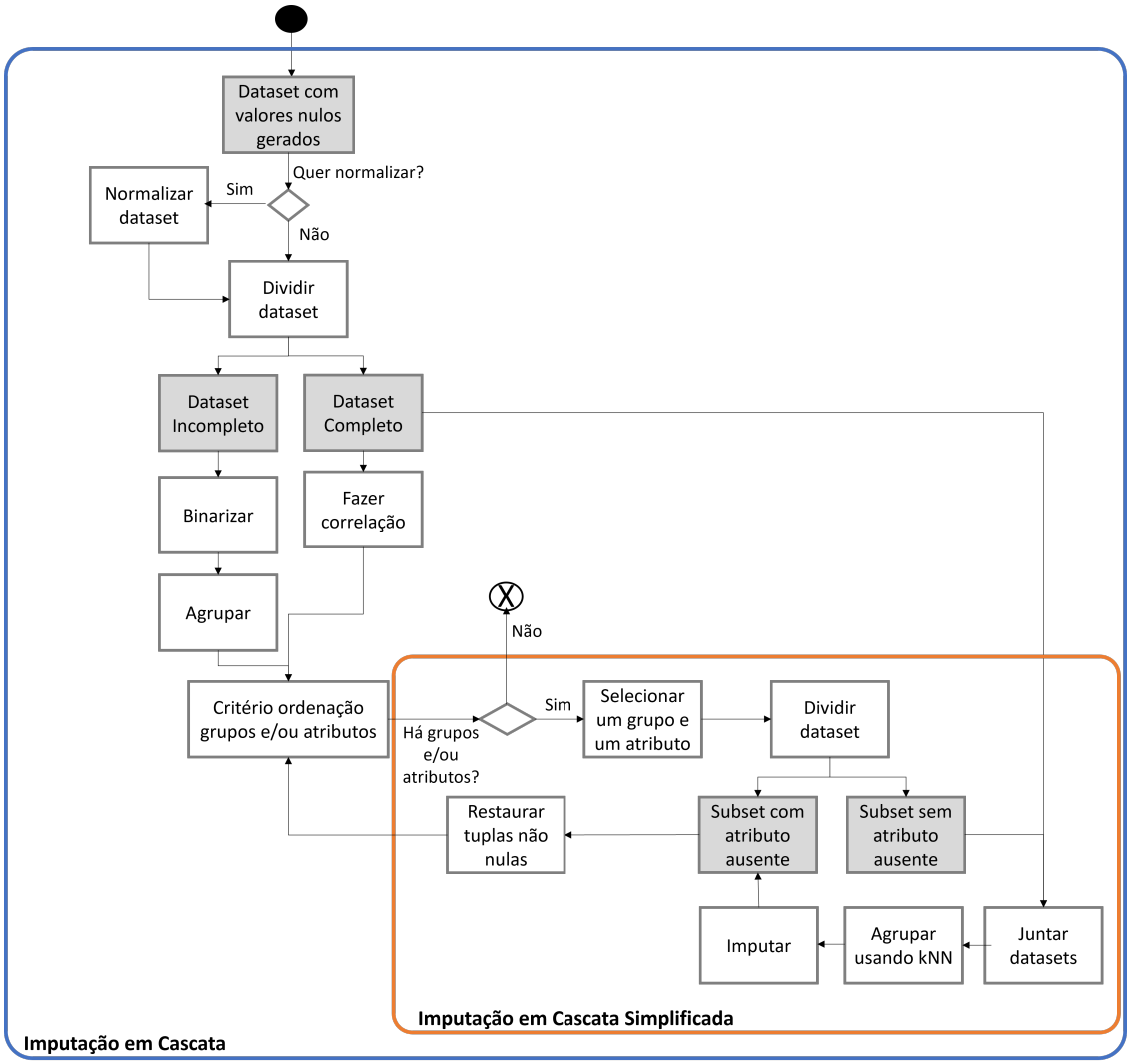


Figura IV.2: Processo completo da Imputação em Cascata (Fonte: Adaptação da proposta de Ferlin [2008]))

No presente estudo, a abordagem que Ferlin [2008] inicialmente descreveu como “Imputação Sequencial com Realimentação” foi renomeada como “Imputação em Cascata Simplificada” (Seção III.2.4). Esta metodologia adota o princípio da reutilização de valores, em que os valores imputados são então reintegrados incrementalmente ao conjunto principal, permitindo que cada etapa subsequente de imputação aproveite as informações adicionadas nas fases anteriores. A decisão de avaliar a “Imputação em Cascata Simplificada” de forma separada surgiu após uma análise preliminar dos resultados das imputações, na qual foi constatado que não existiam diferenças significativas entre

os algoritmos de agrupamento utilizados na "Imputação em Cascata".

A imputação univariada, realizada através da média e da mediana, servirá como método de comparação (*baseline*) para avaliar a eficácia das outras abordagens. Essa técnica será usada como base para determinar se as metodologias mais complexas oferecem vantagens significativas; caso o erro da imputação univariada seja inferior, indica que as metodologias adicionais podem não proporcionar benefícios consideráveis.

Optamos por não realizar a normalização dos dados antes da imputação, ao contrário do procedimento adotado por Ferlin [2008] (Seção III.2.1). Como consideramos a presença de diferentes graus de *outliers* nas bases de dados analisadas, decidimos não normalizar os dados, pois poderíamos comprometer a qualidade da imputação ao alterar as características originais dos dados.

Com base nos detalhes fornecidos nesta seção, é possível visualizar na Figura IV.2 o fluxo de trabalho completo do estudo.

IV.1.1 Formalização da abordagem proposta

Nesta seção, formalizamos o processo de imputação em cascata proposto nesta dissertação, buscando torná-lo abstrato, preciso e estabelecendo um modelo rigoroso e independente de implementações específicas de software.

O processo é definido por elementos-chave detalhados a seguir:

1. D: representa o conjunto de dados original, onde d_{ij} denota o elemento na i-ésima linha e j-ésima coluna. O conjunto D deve conter valores ausentes, que são o foco da imputação.

2. $\rightarrow$: operador que define uma ordem de precedência das tarefas. A expressão $X \rightarrow Y$ significa que a tarefa X precede a tarefa Y.

3. B: uma matriz binária que identifica a presença ou ausência de dados em D. Cada entrada de b_{ij} é definida como 1 se d_{ij} for nulo e 0 caso contrário.

4. A_i: algoritmo utilizado no processo da morfologia da ausência. Exemplo: A_1 = rede SOM, A_2 = K-modes.

5. G: os dados são agrupados com base nos padrões de ausência representados em B. Agrupamentos $G = g_1, g_2, ..., g_k$ são formados, onde cada grupo g consiste em registros com padrões semelhantes de dados ausentes.

6. $O(g, v)$: define a ordem na qual os grupos (g) e os atributos (v) dentro dos grupos serão processados para imputação.

7. $\Rightarrow$: operador que define uma ordem de precedência de algoritmos. A expressão $A_i \Rightarrow A_j$ significa que o algoritmo A_i é aplicado antes do algoritmo A_j.

8. I_i: função de imputação específica para preencher os valores ausentes. Exemplo: I_1 = média, I_2 = k-vizinhos mais próximos (k-NN).

9. $P(v, g, D, I_i)$: plano de imputação I_i de um atributo v do grupo g, de uma base de dados D, aplicado sequencialmente.

10. $C(P, D)$: instância de cascata, aonde os valores atualizados em P, da base de dados D, são reintegrados ao conjunto de dados principal, e o processo é repetido para o próximo grupo ou atributo conforme a ordem O.

11. $\epsilon(C(P, D))$: medida do erro na execução da instância C do plano de imputação P.

O conjunto de valores utilizados para imputação - em um plano de imputação - será composto pelos valores da instância que possuir o menor erro médio entre todas as instâncias daquele plano ($\epsilon(P(v)) = \epsilon(C_k(v))$, onde $\epsilon(C_i(v)) < \epsilon(C_k(v)), \forall i \neq k$). Desse modo, podemos referir a imputação em cascata como a aplicação de uma ou mais estratégias no processo de preenchimento de dados ausentes em um atributo v de um banco de dados D.

IV.2 Conjunto de Dados Utilizados

A seleção dos conjuntos de dados para este estudo foi fundamentada no trabalho de Gonçalves [2021], de revisão sistemática da técnica de imputação *hot-deck*. Esse artigo foi escolhido devido à sua investigação científica ampla na área de imputação, além de o autor analisar previamente as bases mais utilizadas no tema. A partir desse documento, verificou-se as bases de dados mais utilizadas possuíam determinadas características, sendo a publicidade irrestrita desses conjuntos a principal exigência, garantindo a possibilidade de replicação dos experimentos. Adicionalmente, duas características fundamentais se apresentam: (i) os conjuntos de dados deveriam estar completos, sem quaisquer dados faltantes; e (ii) deveriam consistir exclusivamente de atributos numéricos. Devido às complexidades na localização de conjuntos de dados que cumprissem todos os critérios estabelecidos, adaptações foram necessárias. Para os conjuntos de dados com mais de um milhão de linhas, a estratégia adotada foi excluir as tuplas que continham valores nulos (*listwise deletion*), representando aproximadamente 3% das observações; mesmo assim, os conjuntos ainda mantiveram mais de dois milhões de registros. Além disso, atributos categóricos foram removidos dos conjuntos de dados, uma vez que o foco do estudo se restringe exclusivamente à imputação em dados numéricos. Essas decisões visam simplificar o conjunto de dados e garantir a consistência da análise, conforme os objetivos específicos deste trabalho.

Durante o processo de seleção, também se considerou a correlação entre os atributos e a presença de *outliers*, fatores que podem influenciar significativamente os resultados da imputação. A

determinação da correlação entre conjuntos de dados envolveu várias etapas: (i) foi calculada a correlação entre os atributos; (ii) atributos com uma correlação de 0,5 ou mais foram marcados como 1, enquanto os demais receberam a marcação 0. A escolha do valor de correlação de 0,5 foi determinada porque representa uma chance de 50% de pertencer a um dos grupos analisados; (iii) identificou-se o máximo de atributos com volume de alta correlação ($\geq$ 0,5); (iv) posteriormente, calculou-se o percentual do valor máximo em relação total de atributos; e (v) percentuais de correlação dos conjuntos de dados foram ordenados e divididos em três grupos: o grupo 1 contendo os dois conjuntos de dados com mais correlação, o grupo 2 com os três *datasets* subsequentes com mais correlação, e o grupo 3 com os demais conjuntos de dados. A Figura IV.3 demonstra o processo da criação das classes de correlação.

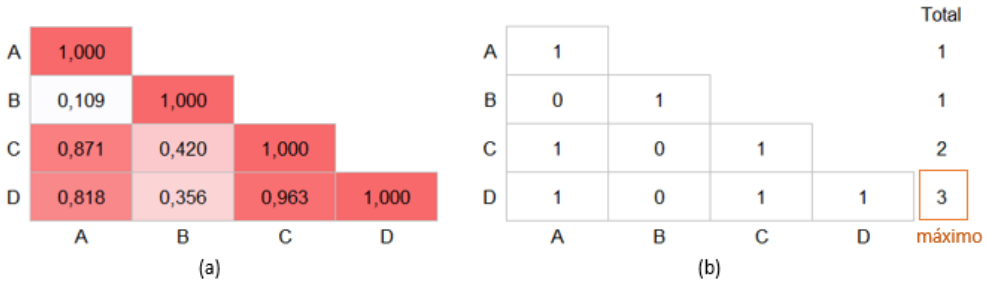
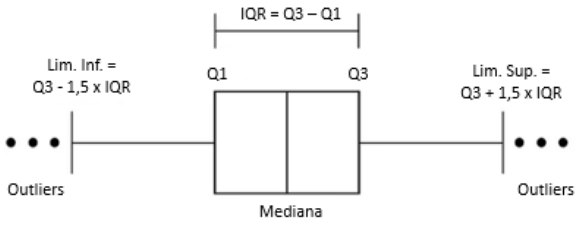


Figura IV.3: Processo de criação das classes de correlação. (a) Matriz de correlação dos atributos. (b) Os atributos com alta correlação ($\geq$ 0,5) foram substituidos por 1 e somados. Com base no valor máximo obtido, o percentual de *outliers* do conjunto de dados em relação ao total de tuplas. (Fonte: Adaptação própria)

A identificação da presença de *outliers* envolveu várias etapas: (i) a identificação de *outliers* foi realizada por meio da análise dos interquartis; (ii) observações identificadas como outliers - ou seja, abaixo do limite inferior ou acima do limite superior - foram marcadas com 1, e as demais com 0; (iii) o percentual de *outliers* foi calculado para cada atributo, considerando o total de tuplas e, em seguida, a média desses percentuais foi determinada para fornecer uma visão geral do conjunto de dados; e (iv) essas médias foram ordenadas e divididas em três grupos iguais: o grupo 1 contendo os dois conjuntos de dados com mais *outliers*, o grupo 2 com os três *datasets* subsequentes com mais *outliers* e o grupo 3 com os demais conjunto de dados. A Figura IV.4 ilustra o processo da criação das classes de *outliers*.

Os conjuntos de dados selecionados foram escolhidos com base na combinação dos grupos de correlação e de *outliers*. Interessante notar que não se encontrou nenhum conjunto de dados que apresentasse simultaneamente alta correlação e presença de *outliers*. A Tabela IV.1 lista os conjuntos de dados escolhidos e suas principais características, ilustrando a aplicação prática desses critérios na seleção dos dados para os testes.



Atributos	# Outliers	# Tuplas Dataset	% Outliers
A	0	200	0,0%
B	5		2,5%
C	1		0,5%
D	0		0,0%
		Média =	0,8%

(b)

Figura IV.4: Processo de criação das classes de *outliers*. (a) Representação do boxplot contendo os Quartis e IQR. (b) O cálculo do percentual de *outliers* para cada atributo é realizado em relação ao total de tuplas. Após isso, calcula-se a média desses percentuais para obter uma medida agregada. (Fonte: Adaptação própria)

Tabela IV.1: Conjunto de dados e suas principais características

Datasets	Local	Classe Correlação	Classe Outlier	Atributos	Tuplas
Wine	UCI	Médio	Baixo	13	178
Pima Indians Diabetes	OpenML	Baixo	Baixo	8	768
Boston Housing	CMU	Alto	Médio	12	506
Abalone	UCI	Alto	Baixo	7	4.177
Forest Fires	CMU	Baixo	Médio	9	517
MAGIC Gamma	UCI	Médio	Médio	10	19.020
Glass Identification	UCI	Médio	Alto	8	214
Yeast	UCI	Baixo	Alto	7	1.484

IV.3 Parâmetros relativos à ausência dos dados

Para simular a ausência de dados nos conjuntos analisados, foram aplicadas taxas de 10%, 20% e 30% de ausência, seguindo o mesmo procedimento adotado no estudo de Ferlin [2008]. Utilizou-se, entretanto, os mecanismos de ausências MCAR, MNAR e MAR, nos datasets analisados - o que diferencia o presente trabalho frente ao de Ferlin [2008], expandindo-o. Esses percentuais de ausências foram aplicados na base como um todo, e mantiveram a proporcionalidade dentro de cada atributo. Por exemplo, considerando a base *Wine*, que possui 178 registros e 13 atributos, totalizando 2.314 células, a aplicação de uma taxa de 10% resultaria na seleção aleatória de 231 células para serem marcadas como ausentes. Isso resulta em diferentes tuplas apresentando de um a n-1 valores ausentes, sendo n é o número total de atributos. Quando consideramos a distribuição de valores ausentes por atributo, cada um apresentará, em média, 10% de seus valores como faltantes.

IV.4 Parâmetros do k-NN

No desenvolvimento deste estudo, optamos por aderir à metodologia utilizada por Ferlin [2008], aplicando os mesmos valores de k: $k = 1$, 3, 5 e 10, mantendo a proposta do trabalho original e, com isso, facilitando a comparação direta dos resultados. Neste estudo, quando o valor de k for maior

que um, a média dos valores dos atributos preenchidos dos vizinhos mais próximos é empregada para imputação.

Diferentemente do trabalho de Ferlin [2008], que explorou tanto a distância Euclidiana quanto a Mahalanobis, neste trabalho optou-se exclusivamente pela utilização da distância Euclidiana. A escolha foi motivada pelo desejo de simplificar o estudo, permitindo focar em outra contribuição deste trabalho que é a imputação em cascata simplificada. Ademais, a utilização da distância Euclidiana é reconhecida como uma referência padrão.

IV.5 Parâmetros dos Algoritmos da Morfologia da Ausência

Este estudo contribuiu com uma análise comparativa das quatro principais categorias de agrupamento, selecionando um algoritmo representativo para cada categoria que fosse compatível tanto com o processo de agrupamento quanto com a binarização dos dados, conforme proposto por Ferlin [2008]. Existem cinco categorias principais de agrupamento, contudo, uma delas não dispõe de um algoritmo que suporte simultaneamente dados numéricos e categóricos. Portanto, optou-se por utilizar apenas os seguintes algoritmos: (i) *Partitioning-based*; (ii) *Hierarchical-based*; (iii) *Density-based*; e (iv) *Model-based*. Essa abordagem propõe-se a aprofundar e expandir os estudos de Ferlin [2008] sobre imputação em cascata, implementando os experimentos no sistema *Appraisal*, uma plataforma desenvolvida por Soares [2007]. O corrente estudo busca avaliar a eficácia de diferentes estratégias de agrupamento dentro do contexto da imputação em cascata.

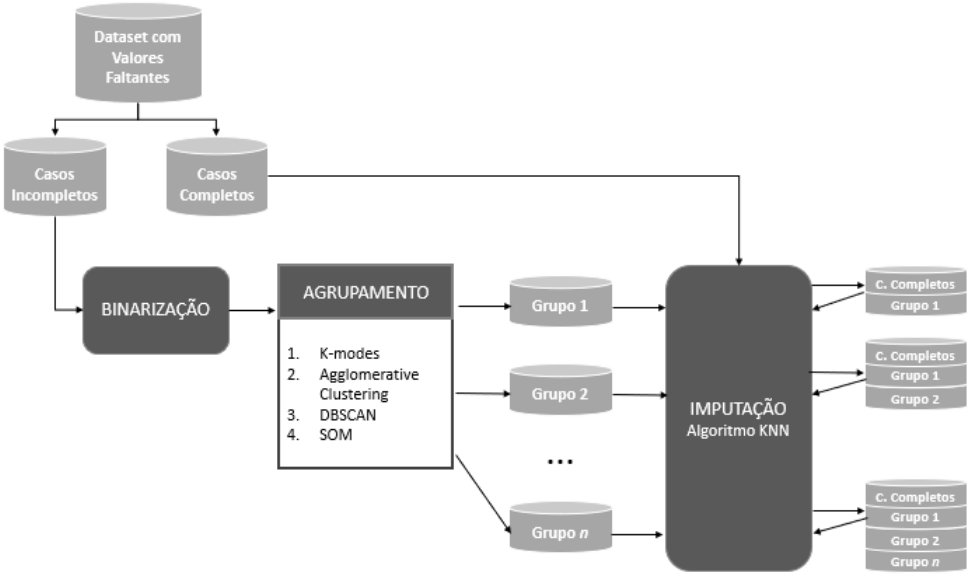


Figura IV.5: Proposta da imputação em cascata de Ferlin [2008] com a aplicação de outras categorias de agrupamento

Os algoritmos escolhidos para este estudo possuem atributos distintos, tornando-os especialmente aptos para abordar variados tipos de dados e necessidades específicas de agrupamento. Essa

seleção meticulosa proporciona um conjunto abrangente de instrumentos analíticos, permitindo uma exploração detalhada e aprofundada dos conjuntos de dados em questão. Ao utilizar essas ferramentas diversificadas, a pesquisa se beneficia da capacidade de cada algoritmo para identificar padrões e agrupamentos de maneira eficaz, adaptando-se às peculiaridades de cada tipo de dado. Na Figura IV.5 apresenta a proposta da aplicação dos algoritmos de agrupamento na imputação em cascata.

Cada algoritmo selecionado possui características únicas que demandam diferentes configurações de parâmetros. Essas configurações são determinadas a priori, visando identificar a combinação que minimiza o erro entre os valores imputados e os valores originais. Nessa seção, serão apresentados os valores de parâmetros que serão testados para cada algoritmo, com o objetivo de alcançar a melhor precisão possível na imputação dos dados.

IV.5.1 K-modes e Agglomerative Clustering

Embora Ferlin [2008] não tenha especificamente trabalhado com os algoritmos *K-modes* e *Agglomerative Clustering*, suas recomendações sobre a escolha dos valores de k para o k-NN foram adotadas neste estudo para esses algoritmos. Portanto, decidiu-se manter os mesmos valores de k para esses dois algoritmos, logo os valores utilizados de k foram 1, 3, 5 e 10.

O *K-modes*, uma adaptação do *K-means* destinada a tratar de atributos categóricos, utiliza a distância de similaridade de *Hamming*. Como foi utilizado o algoritmo disponível em Python, este não permitia a modificação do tipo de distância. Para este estudo, optamos pela inicialização aleatória, que seleciona os centróides iniciais ao acaso, sendo essa a opção *default*.

Para o *Agglomerative Clustering*, além do valor de k, foi necessário definir os parâmetros *affinity* e *linkage*. O parâmetro *affinity* determina a métrica de distância usada para calcular a ligação entre os pontos, e escolhemos a distância de *Gower* para este propósito pois já está implementada em um pacote do Python. Para o parâmetro *linkage*, que define o critério de ligação utilizado, optamos pela ligação *complete*, que calcula as distâncias máximas entre todas as observações dos dois conjuntos comparados.

IV.5.2 DBSCAN

O DBSCAN é um algoritmo que agrupa dados com base na densidade espacial, requerendo uma métrica de distância para calcular a matriz de distância entre as instâncias. Para este estudo, optamos pela distância de *Gower*, adequada para dados binarizados.

No DBSCAN, os parâmetros cruciais incluem *epsilon* e *minimum sample*. O *epsilon* define a distância máxima entre duas amostras para considerá-las como parte da mesma vizinhança, e para nossa aplicação, mantivemos o valor padrão do algoritmo de 0,5. O *minimum sample* determina o número mínimo de amostras necessárias em uma vizinhança para que um ponto seja considerado

um centróide e, para este trabalho, testamos os valores de *minimum sample* de 5, 10, 15, 20, 35 e 50 para explorar diferentes densidades de agrupamento. Essa escolha foi motivada pela necessidade de otimizar a análise para *datasets* com variados tamanhos e características, permitindo uma segmentação eficaz tanto em conjuntos de dados densos quanto esparsos.

IV.5.3 Redes *Self-Organizing Map* (SOM)

Este estudo dá continuidade ao trabalho desenvolvido por Ferlin [2008], esforçando-se para manter a fidelidade em relação aos parâmetros utilizados no algoritmo de agrupamento para possibilitar uma comparação direta entre os resultados. Contudo, a rede neural Self-Organizing Map (SOM) utilizada por Ferlin [2008] apresenta variações nos parâmetros em comparação com a versão implementada em Python empregada neste projeto; então, nesse caso, optamos por usar os valores *default* do algoritmo.

No algoritmo SOM implementado em Python, cinco parâmetros principais precisam ser definidos antecipadamente: dimensões de x e y, sigma, taxa de aprendizado (*learning rate*) e número máximo de iterações.

O parâmetro sigma regula a extensão da vizinhança de um neurônio, isto é, determina a distância máxima na qual uma amostra é considerada vizinha do neurônio. Uma vez que este parâmetro não foi especificado nas configurações adotadas por Ferlin [2008], optou-se pelo valor padrão do algoritmo, que é de 0,5. O *learning rate*, que ajusta o tamanho do passo em cada iteração em direção ao mínimo de uma função de perda, foi mantido em 0,7, conforme utilizado por Ferlin [2008].

Quanto ao parâmetro de número máximo de iterações, definimos os valores de 50, 100 e 500. Essa escolha foi feita considerando os tamanhos das bases de dados a serem simuladas, uma vez que Ferlin [2008] não fornece essa informação e não há um consenso na literatura sobre o valor ótimo para esse parâmetro.

No estudo de Ferlin [2008], diversas configurações de topologia foram exploradas para ajustar o algoritmo SOM às características dos conjuntos de dados utilizados. As dimensões x e y do mapa SOM definem sua topologia, sendo ajustadas de acordo com as seguintes estratégias:

- **Número de Atributos:** para cada base de dados, o mapa possui $2k$ nós, onde k é o número de atributos do conjunto de dados;
- **Número de Tuplas:** a topologia é configurada para ter um nó correspondente a cada caso existente na base de dados;
- **Número de Células Ausentes:** cada célula ausente nos dados é representada por um nó no mapa;

- **Número de Tuplas com Valores Ausentes:** esse parâmetro ajusta o mapa para ter um nó por cada tupla que apresenta pelo menos um valor ausente, o que ajuda na identificação de padrões comuns entre tuplas incompletas; e
- **Média de Atributos Ausentes 5:** a proporção de atributos ausentes em relação ao total de atributos é utilizada para dimensionar o mapa.

Apesar de Ferlin [2008] ter projetado várias configurações de topologia para explorar diferentes perspectivas dos dados ausentes, muitas dessas configurações mostraram-se impraticáveis para este estudo, devido ao uso de conjuntos de dados substancialmente maiores. Por essa razão, a topologia baseada no número de atributos foi a única utilizada neste projeto. Contudo, todas as combinações possíveis de dimensões x e y foram exploradas, como, por exemplo, para a base *Wine*, que possui 13 atributos, as combinações consideradas incluíram 1 x 26, 26 x 1, 2 x 13, e 13 x 2.

Alguns outros parâmetros definidos por Ferlin [2008] são padrão no algoritmo SOM utilizado neste trabalho, como a inicialização aleatória dos protótipos e a determinação da vizinhança. No caso da determinação da vizinhança, implementada com uma distribuição Gaussiana, é a função que influencia a ponderação da proximidade entre posições no mapa.

IV.6 Resumo da metodologia a ser aplicada

Este capítulo delineia a metodologia adotada nesta pesquisa e ressalta as contribuições significativas em comparação com o estudo de Ferlin [2008]. Este último estabeleceu a técnica de imputação em cascata, fundamentada na reutilização de valores. Inicialmente, registros incompletos são segregados do conjunto de dados principal e agrupados conforme características específicas. Esses grupos são subsequentemente reincorporados ao conjunto principal de forma incremental, de modo que cada etapa de imputação possa aproveitar as informações incorporadas nas etapas precedentes.

Este estudo, que se propõe a ser uma continuação do trabalho de Ferlin [2008], não só reavaliou a metodologia de imputação em cascata, mas também introduziu diversas inovações:

1. Enquanto o estudo de Ferlin [2008] se limitou ao uso do mecanismo MCAR, aplicando percentuais de ausência de 10%, 20% e 30%, esta pesquisa expandiu a análise para incluir os três mecanismos de ausência de dados — MAR, MCAR e MNAR — mantendo os mesmos percentuais de valores ausentes.

2. Na seleção dos conjuntos de dados mais empregados na pesquisa, realizou-se uma classificação das bases de acordo com a correlação entre suas variáveis e a presença de outliers, agrupando-as em três categorias: baixa, média e alta. A partir dessa categorização, foram escolhidos oito conjuntos de dados com características distintas, como uma base com baixa correlação

e média presença de outliers. Vale ressaltar que, para os conjuntos de dados selecionados, não foi encontrada nenhum que apresentasse simultaneamente alta correlação e presença de outliers.

3. Foi conduzida uma análise comparativa envolvendo as cinco principais categorias de agrupamento: *Partitioning-based*, *Hierarchical-based*, *Density-based* e *Model-based*. Para cada categoria, selecionou-se um algoritmo representativo apto para trabalhar com dados binários e/ou categorizados.

4. Desenvolveu-se a metodologia denominada Cascata Simplificada, uma adaptação do modelo original de Ferlin [2008], que omite a segmentação inicial baseada no padrão de ausência (morfologia da ausência).

5. Foi feita a formalização da imputação em cascata buscando torná-la abstrata e independente de implementações específicas de software.

Capítulo V Análise dos Resultados

V.1 Descrição dos Experimentos

Para cada conjunto de dados utilizado neste estudo, valores ausentes foram sinteticamente introduzidos, utilizando os três mecanismos de ausência conhecidos: MAR, MCAR e MNAR. Para cada um desses mecanismos, foram aplicados percentuais de ausência de 10%, 20% e 30%. Esta etapa é crucial para uma compreensão adequada dos mecanismos que conduzem à ausência de dados, permitindo assim uma interpretação correta dos resultados das análises. Com esses *datasets* preparados, procedeu-se à aplicação das metodologias de Imputação em Cascata [Ferlin, 2008] e a Imputação em Cascata Simplificada.

Na metodologia da Imputação em Cascata [Ferlin, 2008], o conjunto de dados com valores ausentes sinteticamente gerados é dividido em subconjuntos completo e incompleto. No subconjunto completo, realiza-se a análise de correlação dos atributos, a qual é utilizada para definir o critério de ordenação dos atributos (Seção III.2.3). No subconjunto incompleto, os dados são binarizados, e em seguida, são aplicados os algoritmos K-modes, *Agglomerative Clustering*, DBSCAN e SOM, onde os grupos formados por esses algoritmos são utilizados no critério de ordenação dos segmentos (Seção III.2.3). Posteriormente, os subconjuntos completo e incompleto são reunificados.

Na segunda etapa da metodologia da Imputação em Cascata, seleciona-se um grupo e um atributo específico para iniciar a imputação. O conjunto de dados é então segregado em subconjuntos com e sem o atributo faltante. O conjunto de dados completo e o subconjunto sem atributos faltantes são empregados no agrupamento utilizando o método k-NN. Identificam-se as tuplas semelhantes e a média dos atributos dessas tuplas é utilizada para imputar os valores no subconjunto com atributos faltantes. Após a imputação, as tuplas preenchidas são reintegradas ao conjunto de dados principal e reutilizadas na imputação do atributo subsequente. Este processo é repetido até que não restem mais segmentos ou atributos com valores ausentes a serem tratados.

Na metodologia Imputação em Cascata Simplificada emprega-se o mesmo critério de ordenação de atributos estabelecido na primeira etapa da metodologia Imputação em Cascata. Após selecionar um atributo específico para começar a imputação, o processo segue da mesma forma que a segunda etapa da metodologia da Imputação em Cascata.

Em cada simulação feita no desenvolvimento deste trabalho contém a combinação de parâmetros

de critério de ordenação de segmentos e/ou atributos, dos parâmetros exigidos para cada tipo de algoritmos de agrupamento testado (Seção IV.5) e, por fim, dos parâmetros utilizados no k-NN (Seção IV.4). No total foram feitos um total de 335.520 experimentos (Tabela V.1), evidenciando a complexidade envolvida na geração de resultados para este estudo.

Tabela V.1: Volume de simulações por conjunto de dados e metodologias utilizadas no estudo

Datasets	Simulações		
	Imput. Cascata	Imput. Cascata Simplif.	Total
Wine	37.440	180	37.620
Pima Indians Diabetes	37.440	180	37.620
Boston Housing	54.720	180	54.900
Abalone	37.440	180	37.620
Forest Fires	46.080	180	46.260
MAGIC Gamma	46.080	180	46.260
Glass Identification	37.440	180	37.620
Yeast	37.440	180	37.620
Total	334.080	1.440	335.520

Para cada simulação, calculou-se o erro de imputação utilizando as métricas MAE, MAPE e a medida de erro de similaridade, as quais foram empregadas no estudo de Ferlin [2008]. Além disso, foi avaliada a medida de viés de correlação para investigar o impacto da imputação na correlação entre os atributos, visando entender as alterações estruturais que a imputação pode provocar nos dados.

V.2 Descrição dos Dados

Nesta seção, forneceremos uma descrição detalhada dos conjuntos de dados selecionados para este estudo, explorando aspectos como correlação entre atributos, presença de outliers e outras características relevantes. Esta análise visa esclarecer a composição e a dinâmica dos dados utilizados nas simulações e experimentos subsequentes.

V.2.1 Wine Dataset

A base de dados *Wine*[3] compila resultados de análises químicas de vinhos originários de uma mesma região da Itália, produzidos por três cultivadores diferentes. Esta base é utilizada para estudar as variações químicas e classificar os vinhos de acordo com o cultivador. A base apresenta informações sobre 13 componentes químicos de cada vinho, distribuídos em 178 amostras. Cada amostra pertence a uma das três classes, que correspondem aos três diferentes cultivadores. As classes são distribuídas da seguinte forma: 59 amostras para o primeiro tipo de vinho, 71 para o segundo, e 48 para o terceiro. Os atributos contínuos incluem: *Alcohol* (álcool), *Malic acid* (ácido málico), *Ash*

[3]https://archive.ics.uci.edu/dataset/109/wine

(cinzas), *Alcalinity of ash* (alcalinidade das cinzas), *Magnesium* (magnésio), *Total phenols* (fenóis totais), *Flavanoids* (flavonoides), *Nonflavanoid phenols* (fenóis não flavonoides), *Proanthocyanins* (proantocianidinas), *Color intensity* (intensidade da cor), *Hue* (matiz), *OD280/OD315 of diluted wines* (OD280/OD315 de vinhos diluídos) e *Proline* (prolina). A Tabela V.2 detalha outras características desses dados.

Tabela V.2: Descritivo dos atributos da base *Wine*

Atributo	Tipo	Mínimo	Máximo	Média	Desvio Padrão
Alcohol	real	11	14,83	13	0,81
Malicacid	real	0,7	5,8	2,34	1,12
Ash	real	1,4	3,23	2,37	0,27
Alcalinity of ash	real	10,6	30	19,49	3,34
Magnesium	int	70	162	99,74	14,28
Total phenols	real	1	3,88	2,3	0,63
Flavanoids	real	0,3	5,08	2,03	1
Nonflavanoid phenols	real	0,1	0,66	0,36	0,12
Proanthocyanins	real	0,4	3,58	1,59	0,57
Color intensity	real	1,3	13	5,06	2,32
Hue	real	0,5	1,71	0,96	0,23
0D280/0D315 of diluted wines	real	1,3	4	2,61	0,71
Proline	int	278	1680	746,89	314,91

Os valores máximos e mínimos das colunas (Tabela V.2) indicam uma maior amplitude nos atributos *Magnesium* e *Proline*, o que resulta em um desvio padrão elevado. Na Figura V.1 conseguimos observar que há poucos potenciais outliers nos dados, e conseguimos ter uma visão clara da variabilidade dos atributos.

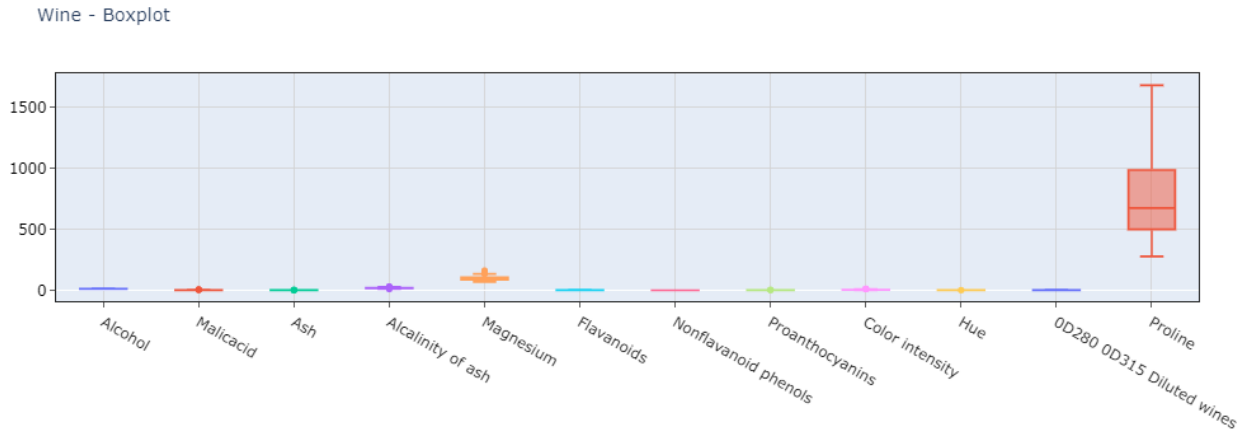


Figura V.1: Boxplot dos atributos da base *Wine*

Como ilustrado na Figura V.2, verifica-se que os atributos *Total phenols* e *Flavanoids* estão altamente correlacionados com diversos outros atributos, enquanto o *Color intensity* se mostra como o menos correlacionado entre eles.

A análise das distribuições dos atributos, através dos histogramas na Figura V.3, revela que

Malicacid, *Magnesium*, *Color intensity* e *Proline* apresentam assimetria à esquerda, indicando que suas médias são inferiores às suas medianas. Adicionalmente, atributos como *Alcohol*, *Total phenols*, *Flavanoids*, e *0D280/0D315 of diluted wines* demonstram comportamento bimodal, exibindo dois picos distintos de frequência.

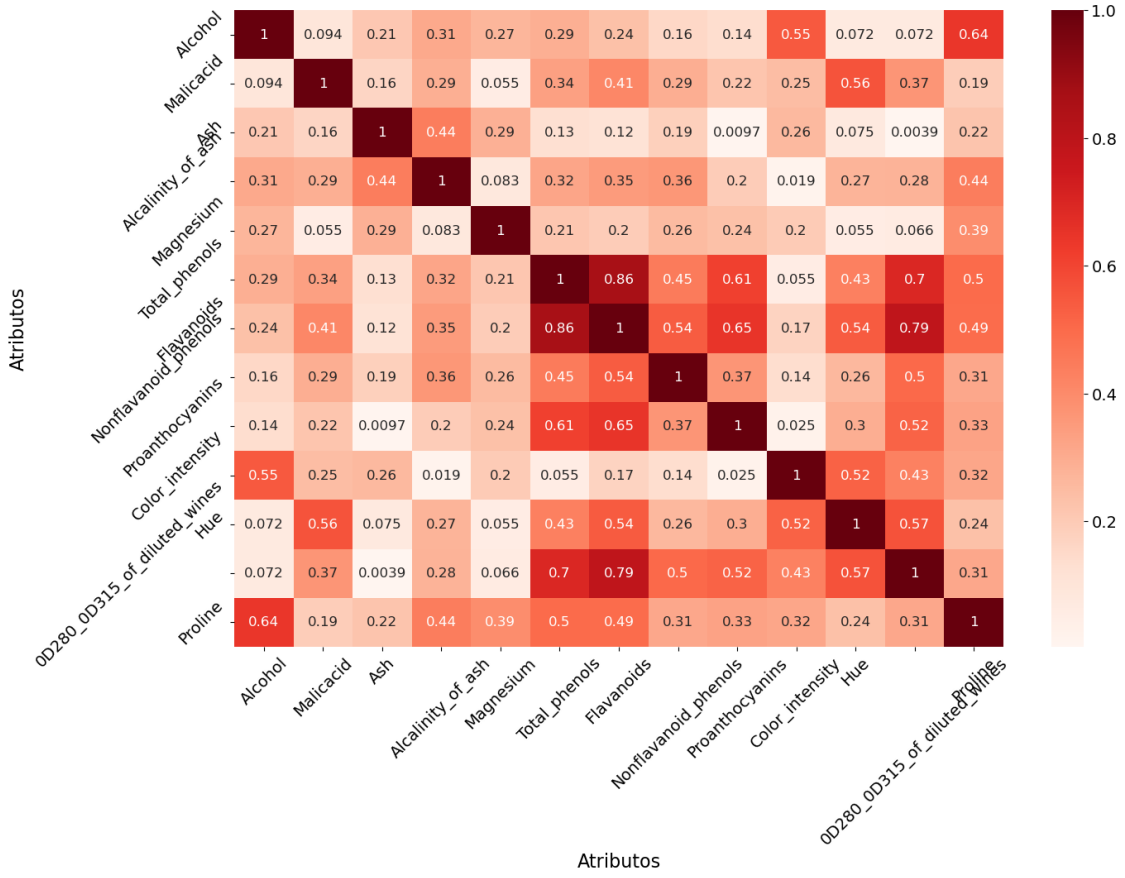


Figura V.2: Matriz de correlação da base *Wine*

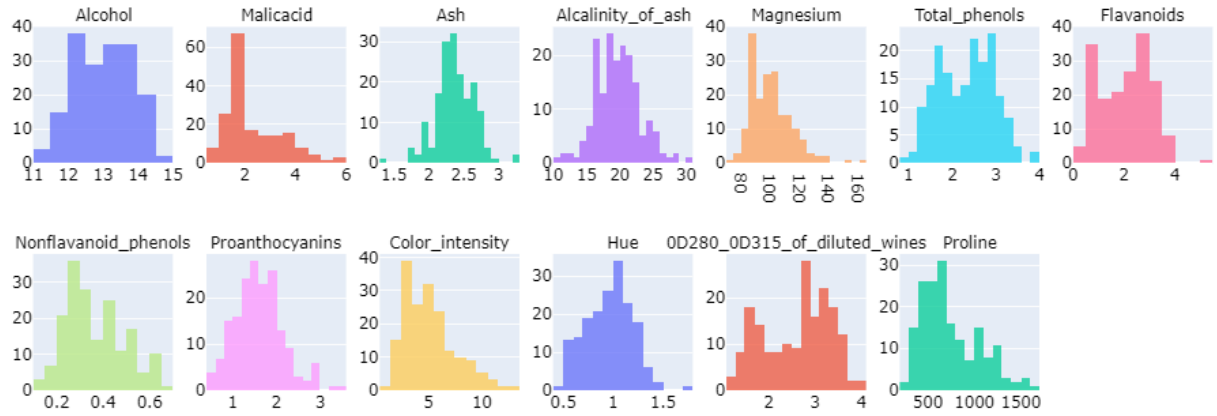


Figura V.3: Histogramas dos atributos da base *Wine*

V.2.2 Pima Indians Diabetes Dataset

A base de dados *Pima Indians Diabetes*[4] reúne informações sobre mulheres da tribo Pima, no Arizona, EUA, todas acima de 21 anos, com o objetivo de estudar a prevalência de diabetes. A base contém 768 registros, dos quais 500 são de mulheres sem diabetes e 268 com diagnóstico positivo para a doença. Os atributos registrados incluem: *Age* (idade da paciente), *Blood Pressure* (pressão sanguínea diastólica), *BMI* (índice de massa corporal), *Glucose* (nível de glicose no sangue duas horas após a ingestão de glicose concentrada de um teste de tolerância a glicose), *Diabetes Pedigree Function* (avaliação das funções hereditárias relativas à diabetes), *Pregnancy* (número de gestacões), *Insulin* (nível de insulina no sangue) e *Skin Thickness* (espessura da pele do tríceps). Essa base é particularmente relevante para estudos que investigam a correlação entre esses fatores e o desenvolvimento de diabetes tipo 2. A Tabela V.3 detalha outras características desses dados.

Tabela V.3: Descritivo dos atributos da base *Pima Indians Diabetes*

Atributo	Tipo	Mínimo	Máximo	Média	Desvio Padrão
Pregnancies	int	0	17	3,85	3,37
Glucose	int	0	199	120,89	31,97
Blood Pressure	int	0	122	69,11	19,36
Skin Thickness	int	0	99	20,54	15,95
Insulin	int	0	846	79,8	115,24
BMI	real	0	67,1	31,99	7,88
Diabetes Pedigree Function	real	0,08	2,42	0,47	0,33
Age	int	21	81	33,24	11,76

A base de dados em análise apresenta uma amplitude razoável entre seus atributos, destacando-se o *Insulin* com um desvio padrão significativo e vários outliers, conforme mostrado na Figura V.4

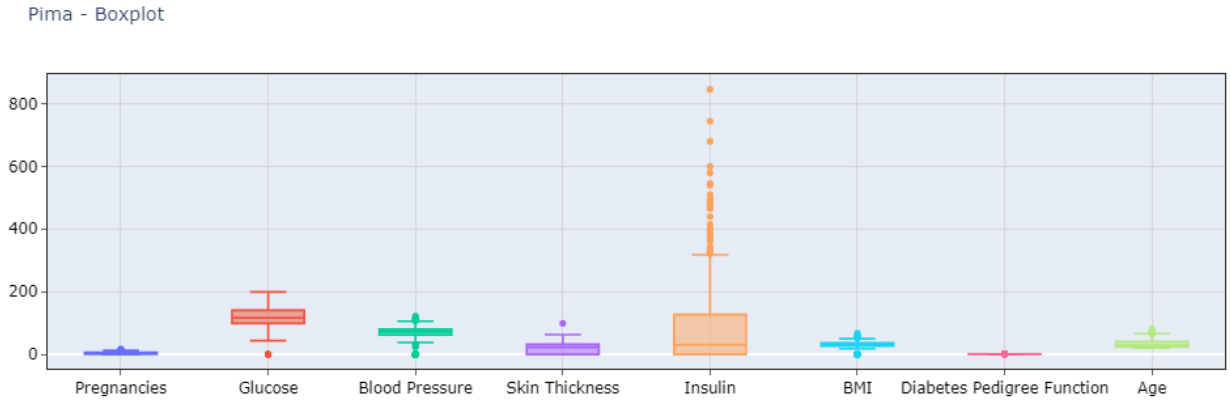


Figura V.4: Boxplot dos atributos da base *Pima Indians Diabetes*

A análise de correlação, ilustrada na Figura V.5, indica a não existência de correlação negativa entre os atributos, ou seja, a variável tende a aumentar quando a outra aumenta. Além da correlação

[4]https://www.openml.org/search?type=data&status=active&id=43582&sort=runs

óbvia entre *Age* e *Pregnancies*, foi observado que existe uma alta correlação entre *SkinThickness*,*BMI* e *Insulin*.

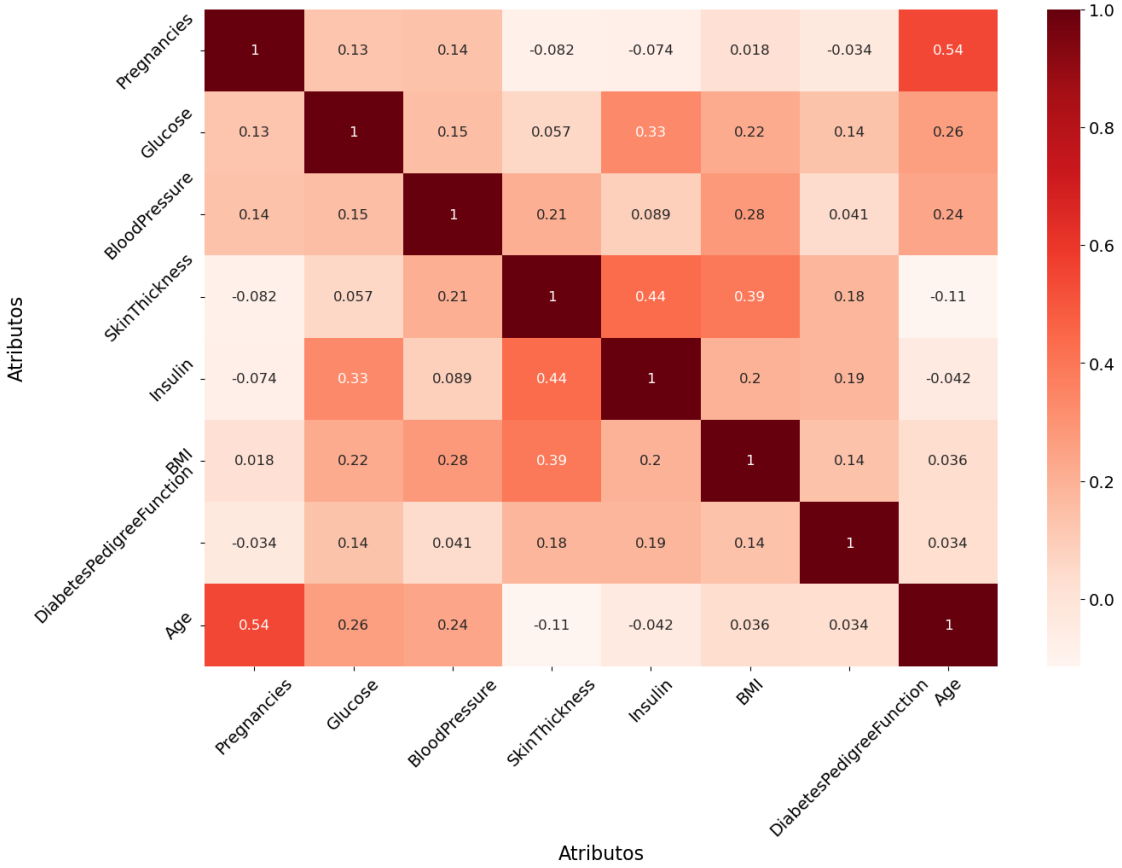


Figura V.5: Matriz de correlação da base *Pima Indians Diabetes*

A distribuição dos atributos (Figura V.6) revela que *SkinThickness* e *Insulin* possuem uma alta incidência de valores zerados. Os atributos *Glucose* e *BloodPressure* exibem um comportamento bimodal, enquanto *Pregnancies*, *DiabetesPedigreeFunction* e *Age* mostram uma distribuição assimétrica à esquerda.

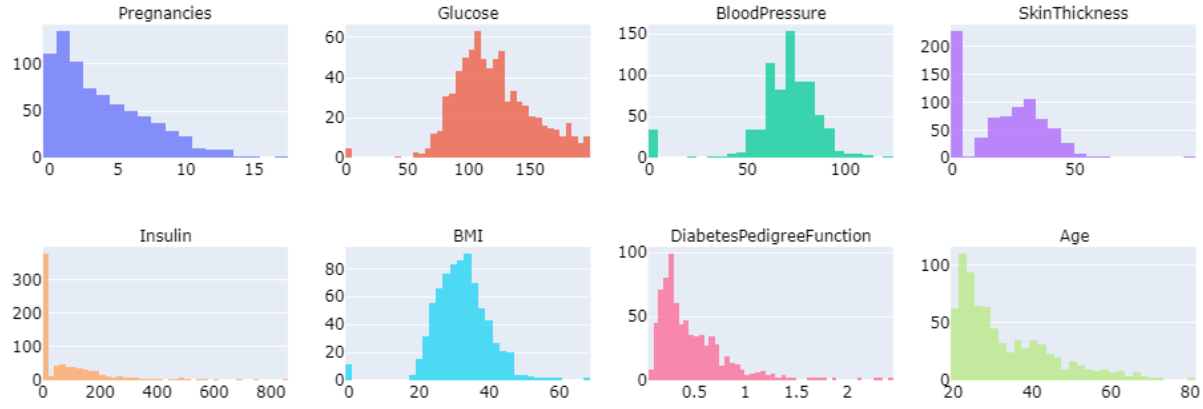


Figura V.6: Histogramas dos atributos da base *Pima Indians Diabetes*

V.2.3 Boston Housing Dataset

O *Boston Housing Dataset*[5] é composto por dados coletados pelo Serviço de Censo dos EUA, que refletem diversas características das habitações e seus moradores na região de Boston, Massachusetts. Este conjunto de dados é frequentemente utilizado em análises de regressão para prever o valor das propriedades com base em diversos atributos. pois possuem características que podem ser utilizadas para modelar e entender as dinâmicas imobiliárias da região, facilitando estudos relacionados ao impacto de variáveis urbanas e ambientais sobre os preços das moradias. As colunas do conjunto de dados incluem: *CRIM* (taxa de criminalidade per capita por cidade), *ZN* (proporção de terrenos residenciais zonados para lotes com mais de 25.000 sq.ft), *INDUS* (proporção de acres não comerciais por cidade), *CHAS* (variável binária onde 1 representa proximidade ao rio Charles (1 se o trato limita o rio; 0 caso contrário), *NOX* (concentração de óxidos nítricos - partes por 10 milhões), *RM* (número médio de quartos por habitação), *AGE* (proporção de unidades ocupadas pelo proprietário construídas antes de 1940), *DIS* (distâncias ponderadas para cinco centros de emprego de Boston), *RAD* (índice de acessibilidade às rodovias radiais), *TAX* (taxa de imposto sobre a propriedade de valor total por US$10.000), *PTRATIO* (proporção aluno-professor por cidade), *B* ($1000(B_k$ - $0.63)^2$, onde B_k é a proporção de pessoas de descendência afro-americana por cidade), *LSTAT* (percentual de população de baixa renda) e *MEDV* (valor médio das casas ocupadas pelos proprietários em US$1000). Para esse trabalho, foram desconsiderados os atributos *CHAS* e *RAD*, pois são variáveis categóricas - as quais não são o foco do estudo. A Tabela V.4 detalha outras características desses dados.

Tabela V.4: Descritivo dos atributos da base *Boston Housing*

Atributo	Tipo	Mínimo	Máximo	Média	Desvio Padrão
CRIM	real	0	89	3,61	8,6
ZN	real	0	100	11,36	23,32
INDUS	real	0,5	27,7	11,14	6,86
NOX	real	0,4	0,9	0,55	0,12
RM	real	3,6	8,8	6,28	0,7
AGE	int	2,9	100	68,57	28,15
DIS	real	1,1	12,1	3,8	2,11
TAX	real	187	711	408,24	168,54
PTRATIO	real	12,6	22	18,46	2,16
B	real	0,3	396,9	356,67	91,29
LSTAT	real	1,7	38	12,65	7,14

Os boxplots analisados (Figura V.7) destacam o atributo *B* devido ao grande número de outliers, seguido pelos atributos *CRIM* e *ZN*. No que diz respeito à correlação entre os atributos (Figura V.8), observa-se que todos apresentam níveis significativos de correlação, tanto positiva quanto negativa.

[5]https://lib.stat.cmu.edu/datasets/boston

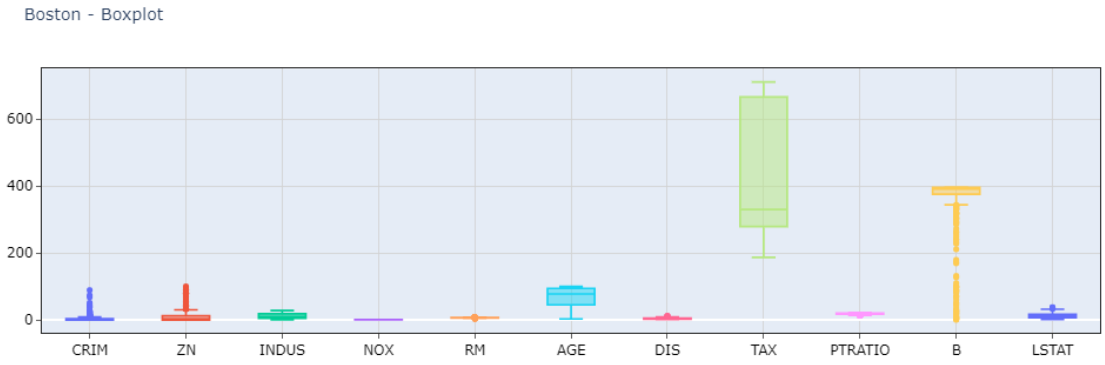


Figura V.7: Boxplot dos atributos da base *Boston Housing*

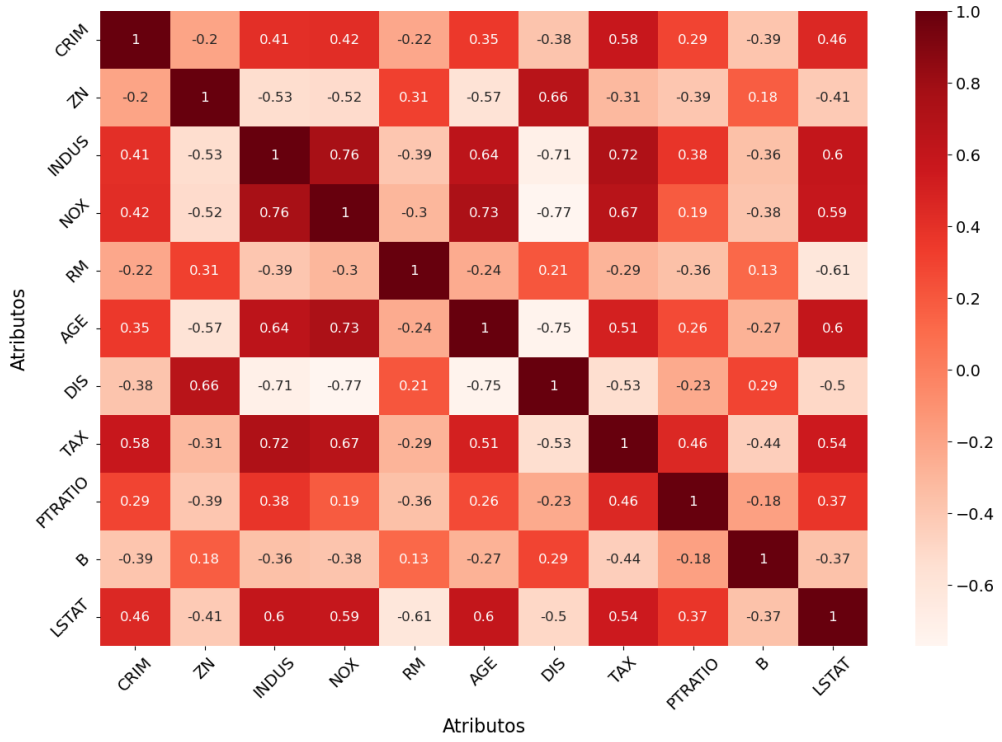


Figura V.8: Matriz de correlação da base *Boston Housing*



Figura V.9: Histogramas dos atributos da base *Boston Housing*

A análise das distribuições dos atributos (Figura V.9) revela comportamentos distintos, com

CRIM e *ZN* exibindo uma prevalência de valores zerados, enquanto *B* mostra uma frequência elevada no valor de 400.

V.2.4 Abalone Dataset

A base de dados *Abalone*[6] reúne características de um grupo de moluscos, chamados de abalone, e tem como objetivo de predizer a idade do mesmo a partir de medições físicas. A base contém 4177 registros, e os atributos registrados incluem: *Sex* (sexo do abalone), *Length* (medida da parte mais comprida da concha), *Diameter* (medida perpendicular ao *Length*), *Height* (medida com a carne na concha), *Whole weight* (abalone inteiro), *Shucked weight* (peso da carne), *Viscera weight* (peso do intestino) e *Shell weight* (peso da concha seca). Para esse trabalho foi desconsiderado o atributo *Sex*, pois é uma variável categórica e não é o foco do estudo. A Tabela V.5 detalha outras características desses dados.

Tabela V.5: Descritivo dos atributos da base *Abalone*

Atributo	Tipo	Mínimo	Máximo	Média	Desvio Padrão
Length	real	0,1	0,8	0,52	0,12
Diameter	real	0,1	0,7	0,41	0,1
Height	real	0	1,1	0,14	0,04
Whole weight	real	0	2,8	0,83	0,49
Shucked weight	real	0	1,5	0,36	0,22
Viscera weight	real	0	0,8	0,18	0,11
Shell weight	real	0	1	0,24	0,14

Na base de dados analisada, o atributo *Whole weight* destaca-se devido ao seu maior desvio padrão, indicando uma variação mais ampla, quando comparado aos outros, conforme detalhado na Tabela V.5.

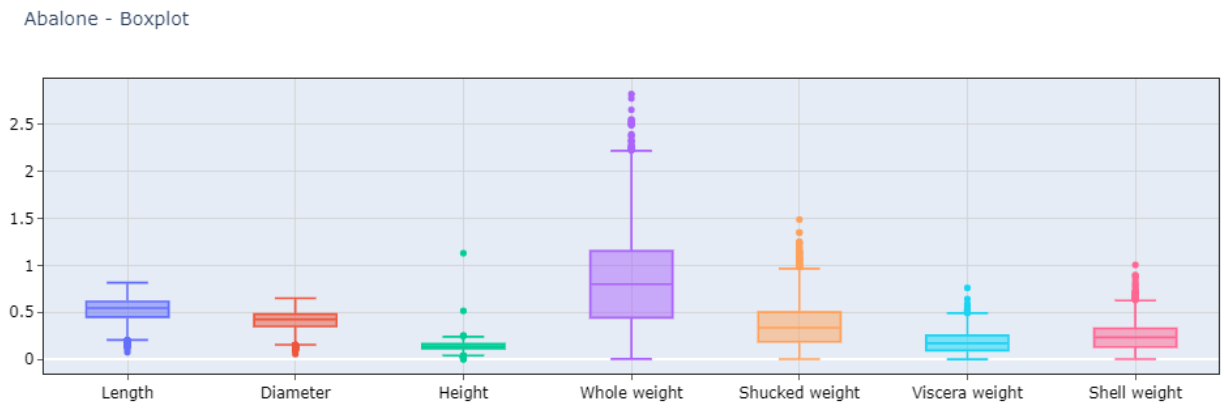


Figura V.10: Boxplot dos atributos da base *Abalone*

Todos os atributos apresentam poucos potenciais outliers nos dados; conseguimos ter uma visão

[6]https://archive.ics.uci.edu/dataset/1/abalone

clara da variabilidade dos atributos (Figura V.10). Em relação à correlação entre os atributos (Figura V.11), todos exibem correlações significativas e exclusivamente positivas.

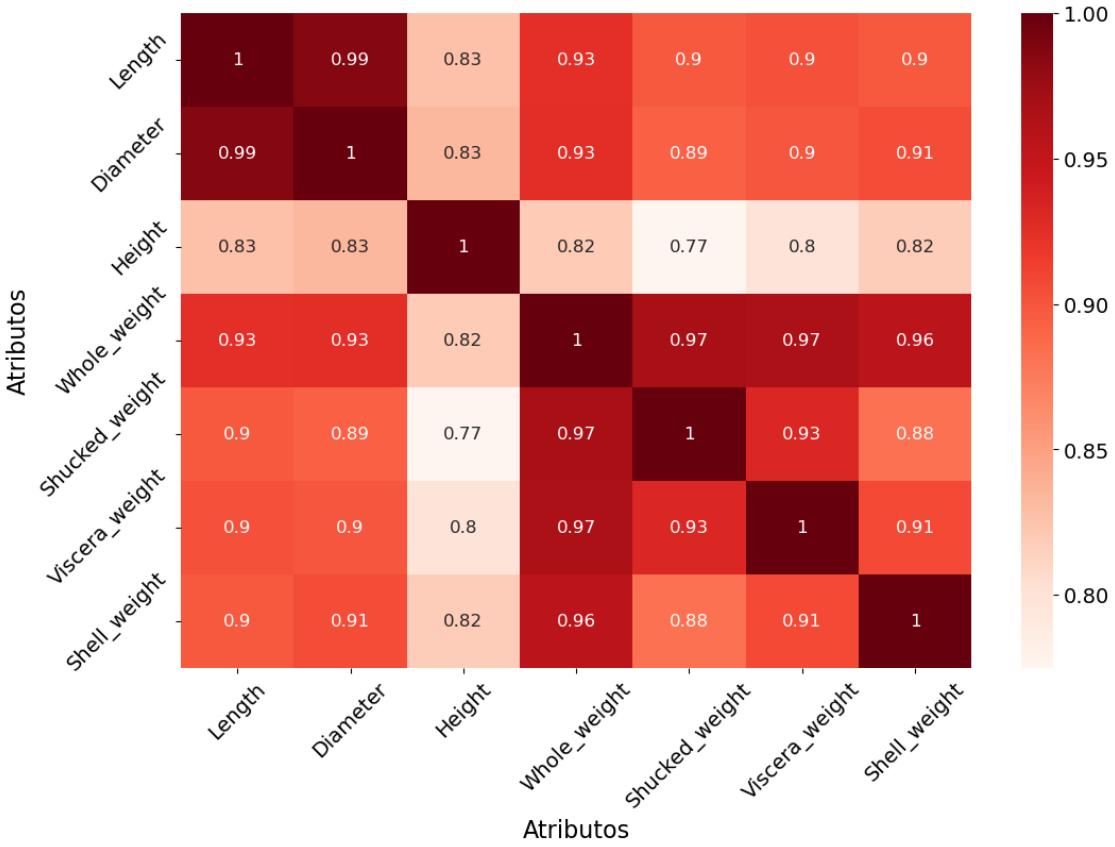


Figura V.11: Matriz de correlação da base *Abalone*

A análise das distribuições dos atributos (Figura V.12) mostra comportamentos similares entre a maioria, exceto *Length* e *Diameter*, que exibem uma assimetria à esquerda.

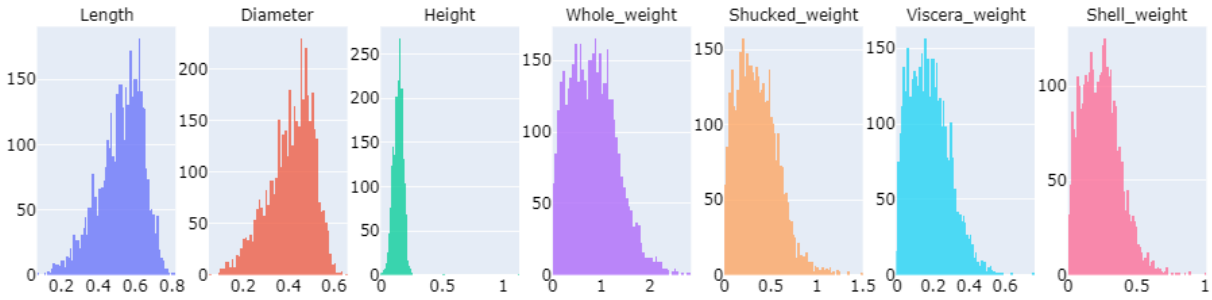


Figura V.12: Histogramas dos atributos da base *Abalone*

V.2.5 Forest Fire Dataset

A base de dados *Forest Fire*[7] reúne características de incêndios florestais do parque natural de Montesinho, do região nordeste de Trás-os-Montes de Portugal. A base contém 517 registros, e os

[7]https://archive.ics.uci.edu/dataset/162/forest+fires

atributos registrados incluem: *X* (coordenada espacial do eixo x no mapa do parque de Montesinho (de 1 a 9)), *Y* (coordenada espacial do eixo y no mapa do parque de Montesinho (de 2 a 9)), *mês* (mês do ano), *dia* (dia da semana), *FFMC* (índice de umidade de combustível do sistema *Forest Fire Weather Index* (FWI)), *DMC* (índice umidade duff do sistema FWI), *DC* (índice seca do sistema FWI), *ISI* (índice de espalhamento inicial do sistema FWI), *temp* (temperatura em graus Celsius), *RH* (umidade relativa em percentual), *vento* (velocidade do vento (em km/h)), *chuva* (chuva externa (em mm/m2)), *área* (área queimada da floresta (em ha)). Para esse trabalho foi desconsiderado o atributo *X*, *Y*, *mês* e *dia*, pois são variáveis categórica e não é o foco do estudo. A Tabela V.6 detalha outras características desses dados.

Tabela V.6: Descritivo dos atributos da base *Forest Fire*

Atributo	Tipo	Mínimo	Máximo	Média	Desvio Padrão
Area	real	0	1090,8	12,85	63,66
FFMC	real	18,7	96,2	90,64	5,52
DMC	real	1,1	291,3	110,87	64,05
DC	real	7,9	860,6	547,94	248,07
ISI	real	0	56,1	9,02	4,56
Temp	real	2,2	33,3	18,89	5,80
RH	inteiro	15	100	44,29	16,32
Wind	real	0,4	9,4	4,01	1,79
Rain	real	0	6,4	0,02	0,29

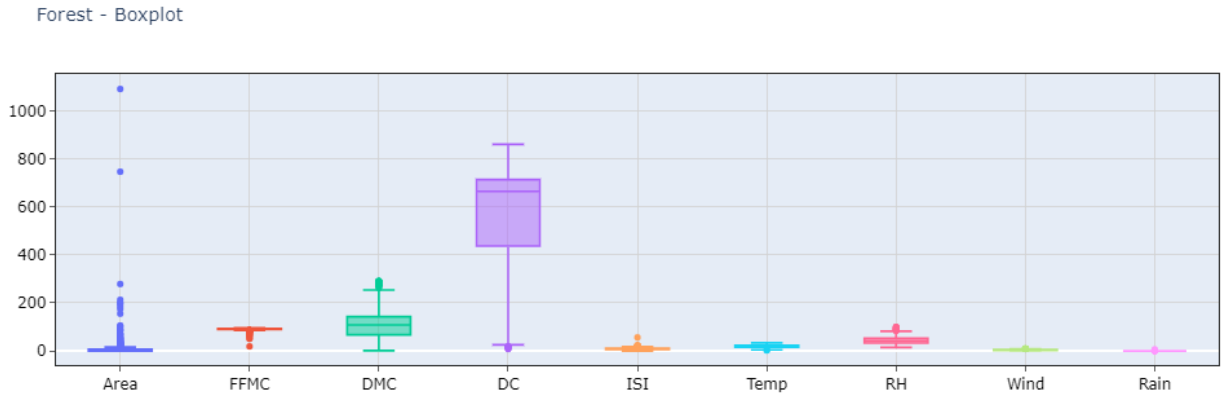


Figura V.13: Boxplot dos atributos da base *Forest Fire*

Os boxplots analisados (Figura V.13) destacam o atributo *Área* devido ao grande número de outliers, seguido pelos atributos *DMC* e *RH*. No que diz respeito à correlação entre os atributos (Figura V.14), observa-se apenas os atributos *Área*, *Wind* e *Rain* apresentam níveis baixos de correlação, tanto positiva quanto negativa.

A análise das distribuições dos atributos (Figura V.15) revela comportamentos distintos, com *Área* e *Rain* exibindo uma prevalência de valores zerados, enquanto *FFMC* mostra uma frequência elevada no valor de 92,9.

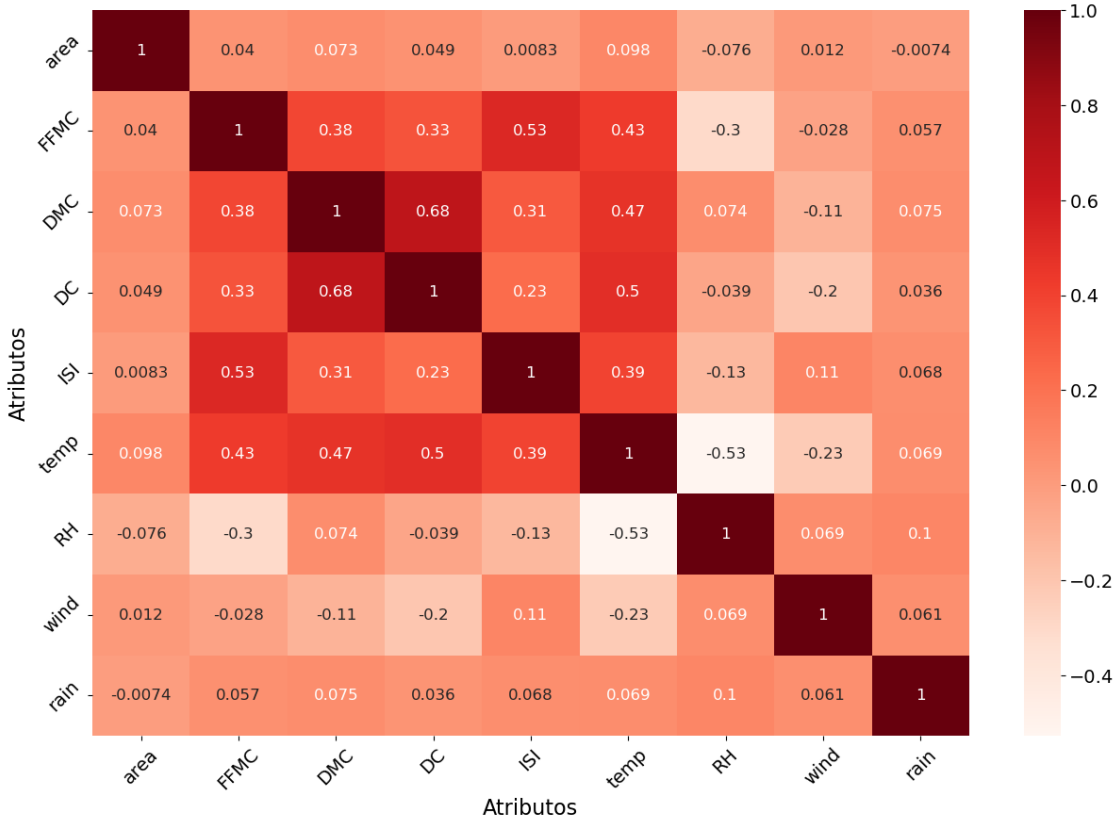


Figura V.14: Matriz de correlação da base *Forest Fire*

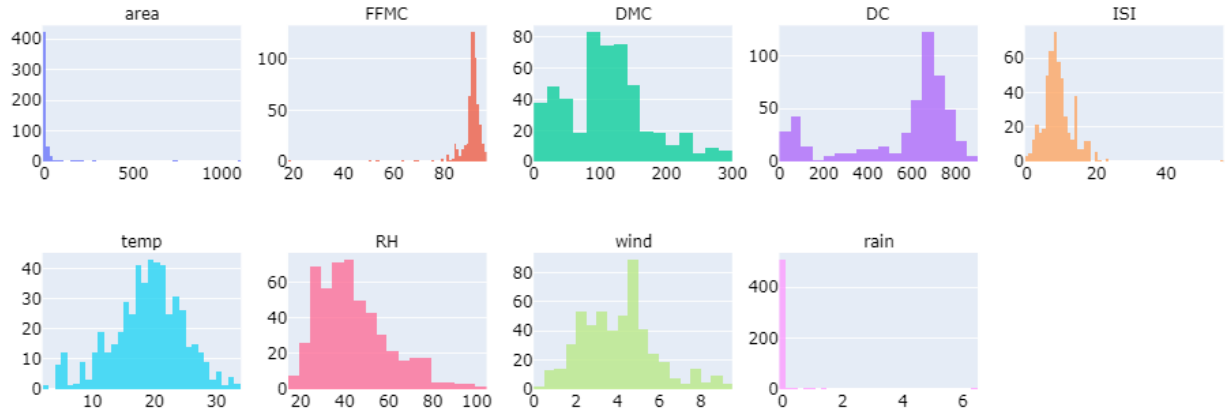


Figura V.15: Histogramas dos atributos da base *Forest Fire*

V.2.6 Magic Gamma Dataset

A base de dados *Magic Gamma*[8] foi gerada por meio de um programa de Monte Carlo, com o objetivo de simular o registro de partículas gama de alta energia em um telescópio Cherenkov atmosférico de superfície, utilizando a técnica de imageamento. A base contém 19.020 registros, e os atributos registrados incluem: *fLength* (eixo maior da elipse (em mm)), *fWidth* (eixo menor da elipse (em mm)), *fSize* (10-log da soma do conteúdo de todos os pixels), *fConc* (proporção da soma

[8]https://archive.ics.uci.edu/dataset/159/magic+gamma+telescope

dos dois pixels mais altos sobre fSize), *fConc1* (proporção do pixel mais alto sobre fSize), *fAsym* (distância do pixel mais alto ao centro, projetada no eixo maior), *fM3Long* (raiz cúbica do terceiro momento ao longo do eixo maior (em mm)), *fM3Trans* (raiz cúbica do terceiro momento ao longo do eixo menor (em mm)), *fAlpha* (ângulo do eixo maior com o vetor até a origem (em graus)), *fDist* (distância da origem ao centro da elipse(em mm)). Para esse trabalho foi desconsiderado o atributo *class* pois é uma variável categórica, além de ser o target desse dataset, e não é o foco do estudo. A Tabela V.7 detalha outras características desses dados.

Tabela V.7: Descritivo dos atributos da base *Glass Identification*

Atributo	Tipo	Mínimo	Máximo	Média	Desvio Padrão
fLength	real	4,3	334,2	53,25	42,36
fWidth	real	0	256,4	22,18	18,35
fSize	real	1,9	5,3	2,82	0,47
fConc	real	0,01	0,9	0,38	0,18
fConc1	real	0,0003	0,7	0,21	0,11
fAsym	real	-457,9	575,2	-4,33	59,20
fM3Long	real	-331,8	238,3	10,54	51,00
fM3Trans	real	-205,9	179,8	0,25	20,83
fAlpha	real	0	90	27,65	26,10
fDist	real	1,3	495,6	193,82	74,73

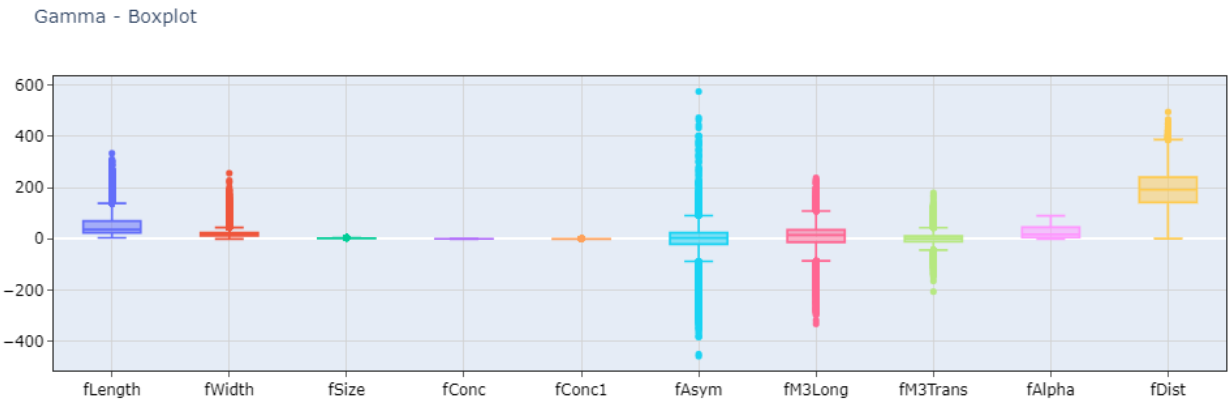


Figura V.16: Boxplot dos atributos da base *Magic Gamma*



Figura V.17: Histogramas dos atributos da base *Magic Gamma*

Os boxplots analisados (Figura V.16) destacam que os atributos *fSize*, *fConc*, *fConc1* e *fAlpha* não possuem outliers. No que diz respeito à correlação entre os atributos (Figura V.18), observa-se que as variáveis *fLength*, *fWidth*, *fSize* e a *fDist* apresentam uma correlação positiva entre si, ou seja, quando uma variável aumenta, as demais também tendem a aumentar. Além disso, *fConc* e *fConc1* exibem uma correlação negativa com os atributos *fLength*, *fWidth* e *fSize*, indicando que, à medida que uma delas aumenta, a outra diminui.

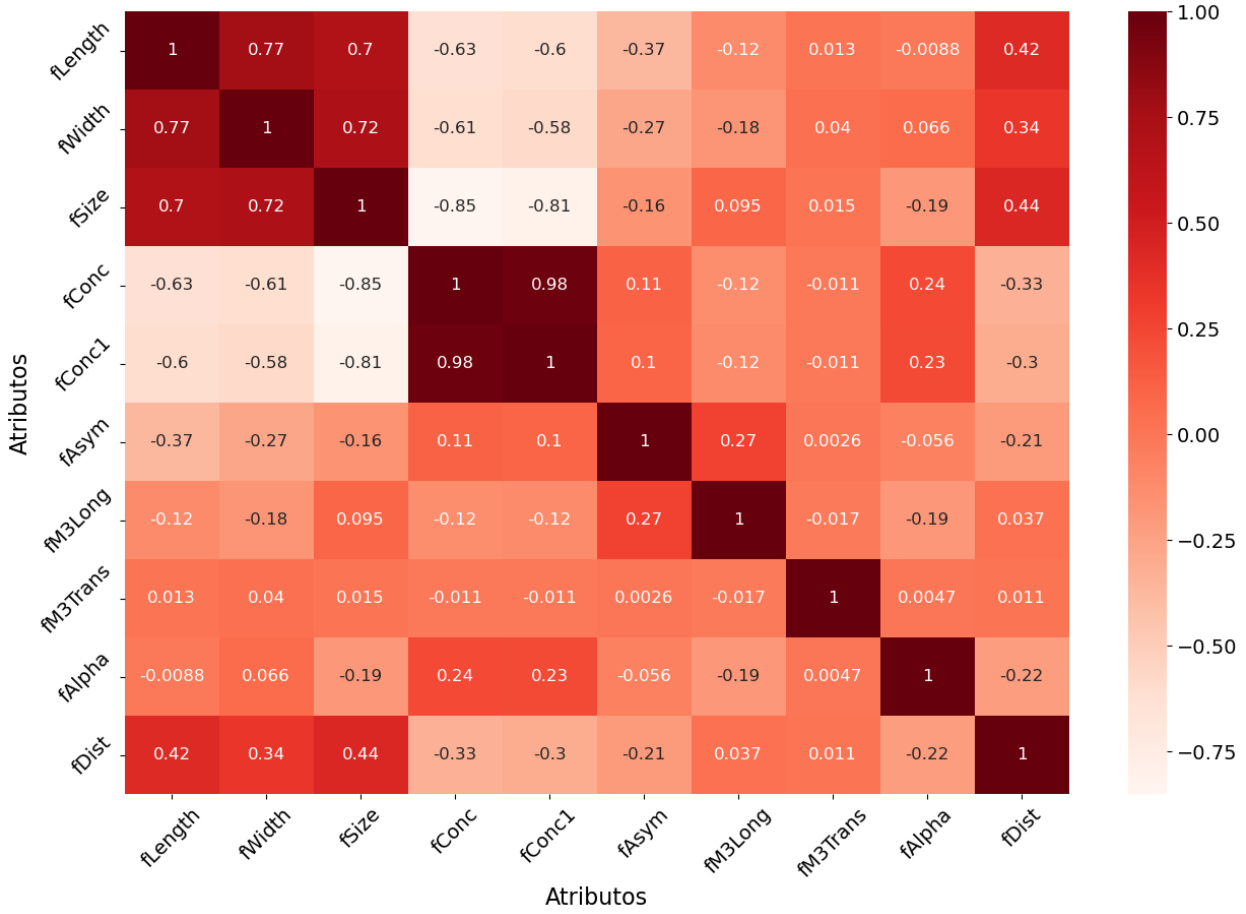


Figura V.18: Matriz de correlação da base *Magic Gamma*

A análise das distribuições dos atributos (Figura V.17) revela que *fLength* e *fWidth* apresentam um comportamento semelhante, assim como *fSize*, que se correlaciona com os atributos *fConc* e *fConc1*. Além disso, os atributos *fM3Long* e *fM3Trans* exibem uma distribuição bimodal.

V.2.7 Glass Identification Dataset

A base de dados *Glass Identification*[9] reúne características sobre tipos de vidros. A base contém 214 registros, e os atributos registrados incluem: *Id number* (número de identificação do vidro), *RI* (índice de refração), *Na* (sódio (percentagem em peso no óxido correspondente)), *Mg* (magnésio (percentagem em peso no óxido correspondente)), *Al* (alumínio (percentagem em peso no

[9]https://archive.ics.uci.edu/dataset/42/glass+identification

óxido correspondente)), *Si* (silício (percentagem em peso no óxido correspondente)), *K* (potássio (percentagem em peso no óxido correspondente)), *Ca* (cálcio (percentagem em peso no óxido correspondente)), *Ba* (bário (percentagem em peso no óxido correspondente)), *Fe* (ferro (percentagem em peso no óxido correspondente)), *Type of glass* (tipo de vidros). Para esse trabalho foi desconsiderado o atributo *Id number* e *Type of glass*, pois são variáveis categórica e não é o foco do estudo. A Tabela V.8 detalha outras características desses dados.

Tabela V.8: Descritivo dos atributos da base *Glass Identification*

Atributo	Tipo	Mínimo	Máximo	Média	Desvio Padrão
RI	real	1,5	1,5	1,52	0,003
Na	real	10,7	17,4	13,41	0,82
Mg	real	0	4,5	2,68	1,44
Al	real	0,3	3,5	1,45	0,50
Si	real	69,8	75,4	72,65	0,77
K	real	0	6,2	0,50	0,65
Ca	real	5,4	16,2	8,96	1,42
Ba	real	0	3,1	0,17	0,50
Fe	real	0	0,5	0,06	0,10

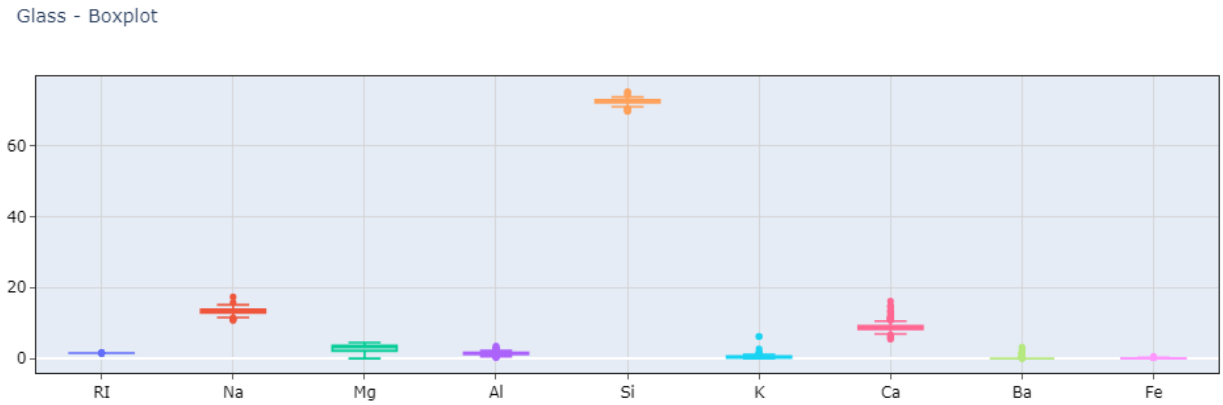


Figura V.19: Boxplot dos atributos da base *Glass Identification*

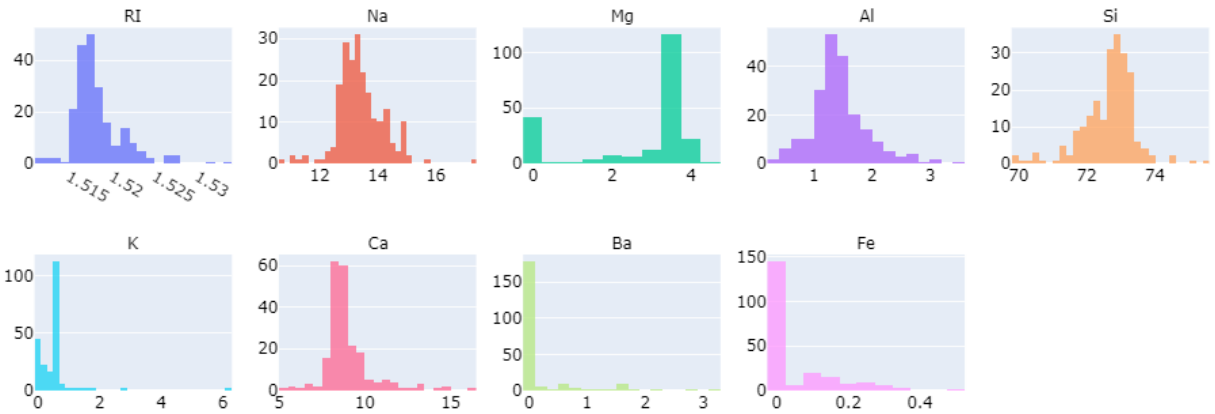


Figura V.20: Histogramas dos atributos da base *Glass Identification*

Os boxplots analisados (Figura V.19) destacam que os atributos *K* e *Ca* possuem um pouco de outliers em relação aos demais atributos. No que diz respeito à correlação entre os atributos (Figura V.21), observa-se que o *RI* e o *Ca* possuem alta correlação positiva.

A análise das distribuições dos atributos (Figura V.20) mostra que *RI* e *Ca* possuem comportamento semelhante, confirmando a alta correlação vista na Figura V.19.

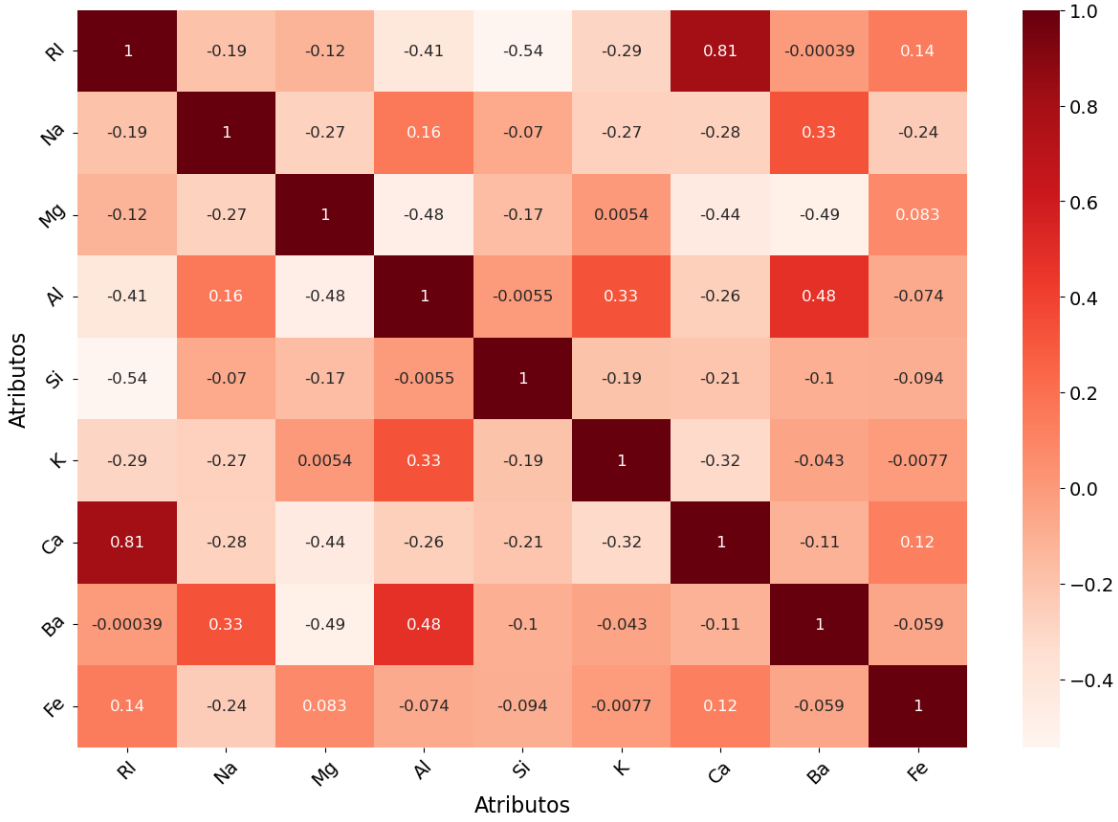


Figura V.21: Matriz de correlação da base *Glass Identification*

V.2.8 Yeast Dataset

A base de dados *Yeast*[10] reúne informações sobre a localização da proteína. A base contém 1.484 registros, e os atributos registrados incluem: *Sequence Name* (número de acesso à base de dados SWISS-PROT), *mcg* (método de McGeoch para reconhecimento de sequência de sinal), *gvh* (método de von Heijne para reconhecimento de sequência de sinal), *alm* (pontuação do programa de predição da região que abrange a membrana ALOM), *com* (pontuação da análise discriminante do conteúdo de aminoácidos da região N-terminal de proteínas mitocondriais e não mitocondriais), *erl* (presença de substring HDEL), *pox* (sinal de direcionamento peroxissomal no terminal C), *vac* (pontuação da análise discriminante do conteúdo de aminoácidos de proteínas vacuolares e extracelulares), *nuc* (pontuação da análise discriminante de sinais de localização nuclear de proteínas nucleares e não nucleares) e *localization site* (localização da proteína). Para esse trabalho foi desconsiderado o

[10]https://archive.ics.uci.edu/dataset/110/yeast

atributo *Sequence Name* e *localization site*, pois são variáveis categórica e não é o foco do estudo. A Tabela V.9 detalha outras características desses dados.

Tabela V.9: Descritivo dos atributos da base *Yeast*

Atributo	Tipo	Mínimo	Máximo	Média	Desvio Padrão
MCG	real	0,11	1	0,5	0,14
GVH	real	0,13	1	0,5	0,12
ALM	real	0,21	1	0,5	0,09
MIT	real	0	1	0,26	0,14
ERL	real	0,5	1	0,5	0,05
POX	real	0	0,83	0,01	0,08
VAC	real	0	0,73	0,5	0,06
NUC	real	0	1	0,28	0,11

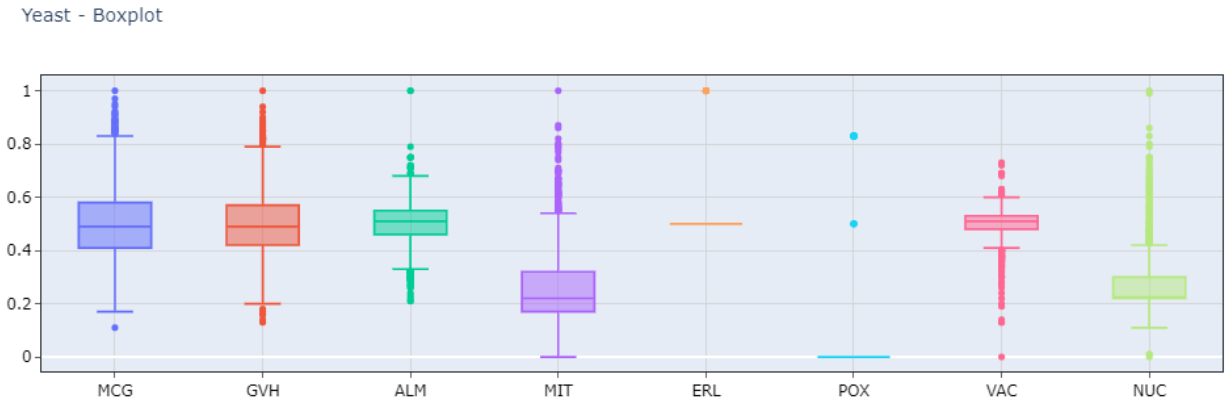


Figura V.22: Boxplot dos atributos da base *Yeast*

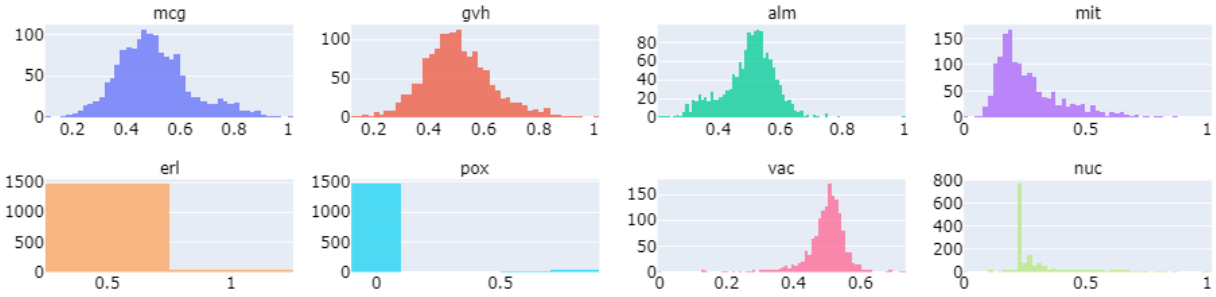


Figura V.23: Histogramas dos atributos da base *Yeast*

Os boxplots analisados (Figura V.22) destacam que todos os atributos possuem grandes volumes de outliers. No que diz respeito à correlação entre os atributos (Figura V.24), observa-se que apenas o *MCG* e o *GVH* são correlacionados.

A distribuição dos atributos (Figura V.23) revela que *POX* possui uma alta incidência de valores zerados. Os atributos *MCG* e o *GVH* possuem comportamentos semelhantes, enquanto *MIT* tem distribuição assimétrica à esquerda e *VAC* com distribuição assimétrica à direita.

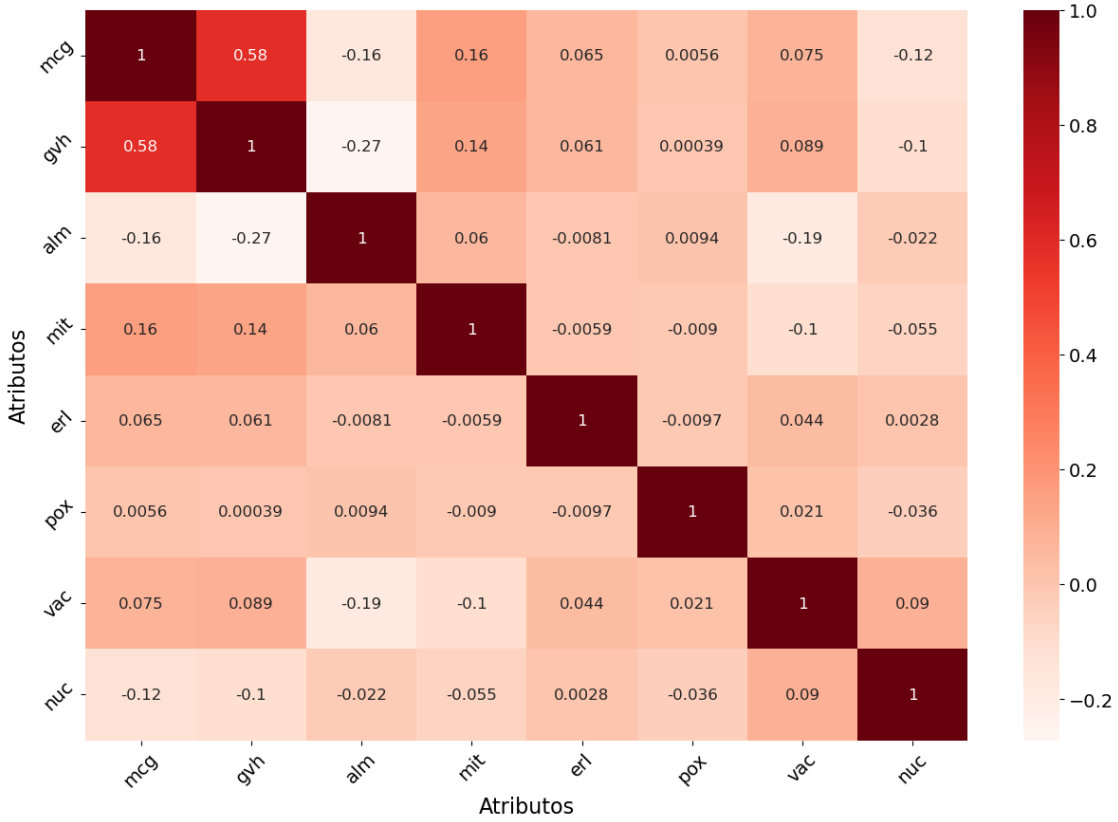


Figura V.24: Matriz de correlação da base *Yeast*

V.3 Condições ambientais dos experimentos

O workflow do trabalho, utilizando linguagem Python, foi desenvolvido em um microcomputador com processador Intel Core i5, com 8 GB de RAM e 24 GB de disco SSD, com o sistema operacional Windows 11 Pro. Entretanto, os testes foram realizados em quatro servidores com as seguintes características:

- **Adhafera:** Intel(R) Core(TM) i7-7740X CPU @ 4.30GHz, 64 GB de RAM e Ubuntu 20.04.6 LTS.
- **Cancri:** Intel(R) Core(TM) i7-14700F @ 5.4 GHz, 32 GB de RAM e Ubuntu 24.04.1 LTS.
- **Capricorni:** Intel(R) Core(TM) i7-7740X CPU @ 4.30GHz, 64 GB de RAM e Ubuntu 22.04.5 LTS.
- **Pegasi:** Intel(R) Core(TM) i7-7740X CPU @ 4.30GHz, 32 GB de RAM e Ubuntu 22.04.4 LTS.

V.4 Análises dos Resultados

Foram realizadas diversas simulações que abrangeram todos os quatro algoritmos de agrupamento, variando alguns parâmetros, critérios de ordenação dos segmentos e atributos, além da imputação em cascata. A eficácia do processo foi avaliada comparando os valores imputados com os dados originais, utilizando o percentual do erro. Adicionalmente, os resultados foram comparados com a imputação por média e por mediana, metodologias que foram utilizadas como referência básica. A técnica de imputação por média é comumente empregada, enquanto a imputação por mediana foi escolhida neste estudo para atenuar o impacto de outliers nas bases de dados.

V.4.1 Comparação do desempenho da Imputação em Cascata em vários Mecanismos de Ausências

Nesta seção, realizaremos um comparativo entre o desempenho das abordagens de imputação em cascata, imputação em cascata simplificada e imputação univariada (média e mediana), conforme descrito na seção IV.1.

Avaliação dos Erros Percentuais na Imputação em Cascata

Embora Ferlin [2008] tenha analisado apenas o erro mínimo percentual, optamos, neste trabalho, por também avaliar o erro médio percentual e investigar se existem variações significativas entre esses dois indicadores.

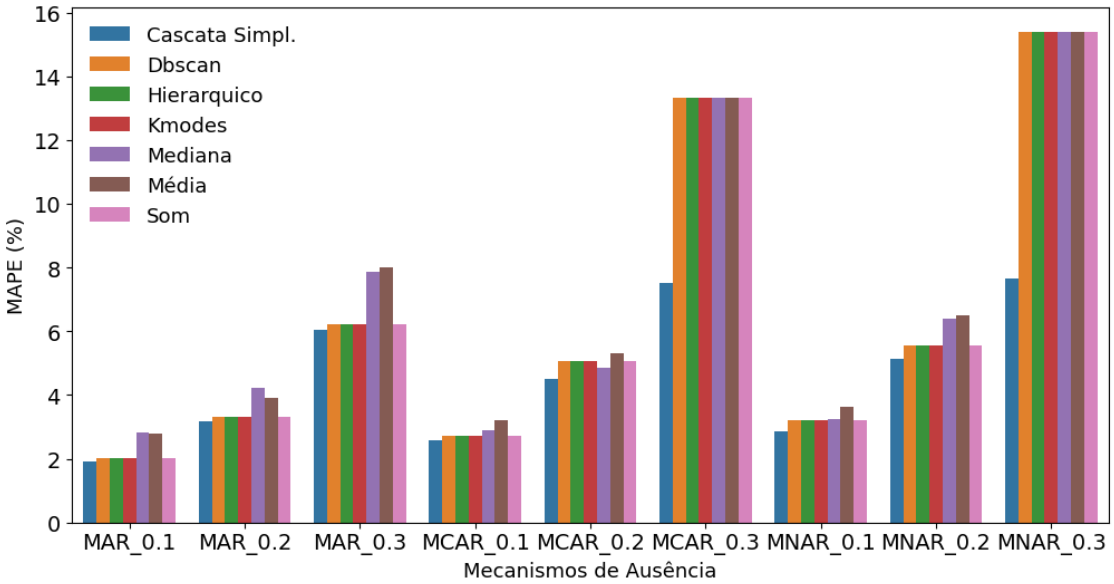


Figura V.25: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Wine*

Ao comparar os erros médios e mínimos de imputação da base *Wine*, observa-se que a diferença

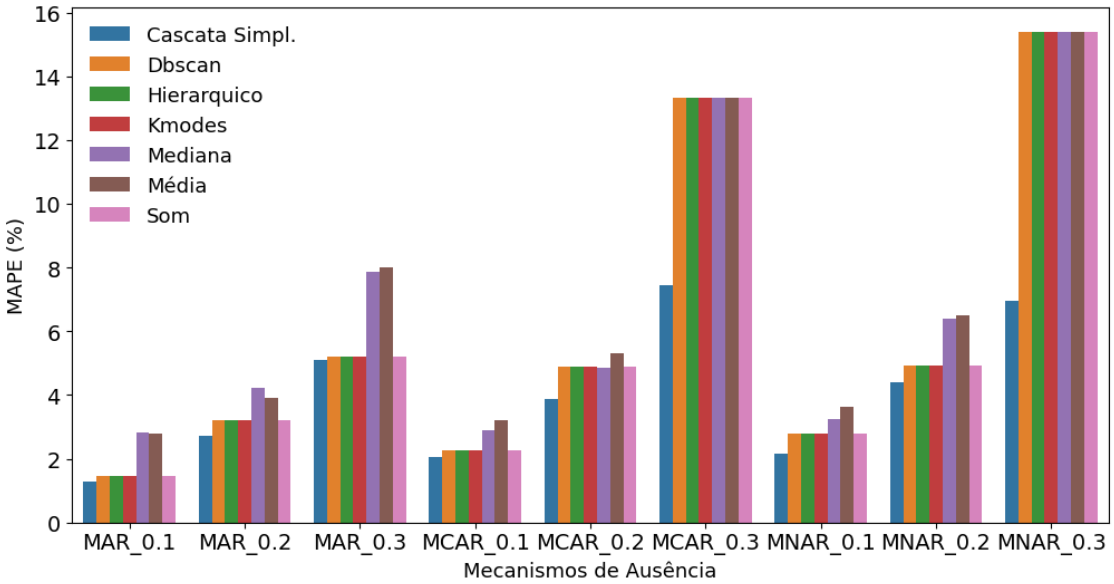


Figura V.26: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Wine*

entre esses indicadores é pouco significativa, indicando que a distribuição dos erros é similar, conforme evidenciado nas Figuras V.25 e V.26. Além disso, com exceção dos mecanismos MCAR 30% e MNAR 30%, todos os métodos de imputação testados apresentaram um desempenho superior ao da imputação pela média ou mediana. Notou-se ainda que, para todos os mecanismos de ausência, o método de imputação em Cascata Simplificada obteve o menor erro de imputação.

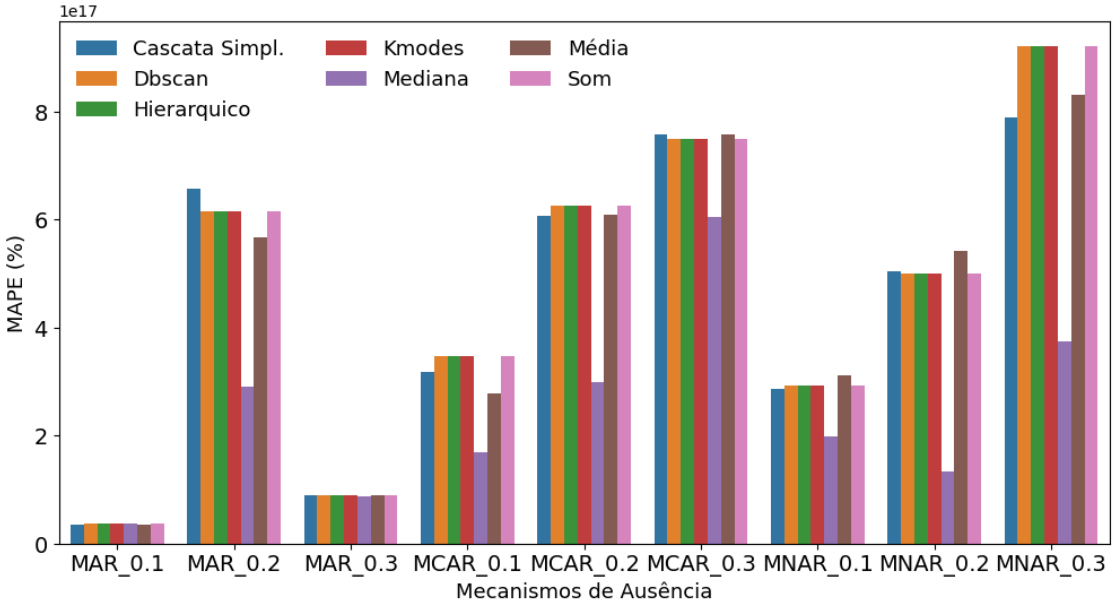


Figura V.27: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Pima*

Na comparação dos erros médios e mínimos de imputação da base *Pima*, ao analisar o erro médio (Figura V.27) nos mecanismos de ausência MAR 20% e MCAR 30%, observa-se um desempenho

inferior do método de imputação em Cascata Simplificada. Por outro lado, ao avaliar o erro mínimo (Figura V.28), este método se destacou como o de melhor desempenho, apresentando o menor erro mínimo. Ademais, verificou-se que a base *Pima* teve um erro de imputação consideravelmente menor, em todos os métodos, nos mecanismos de ausência MAR 10% e MAR 30%. Destaca-se também que a imputação por mediana apresentou um desempenho superior, seguida da Cascata Simplificada, em quase todos os mecanismos de ausência, exceto nos casos de MAR 10% e MAR 30% (Figura V.28).

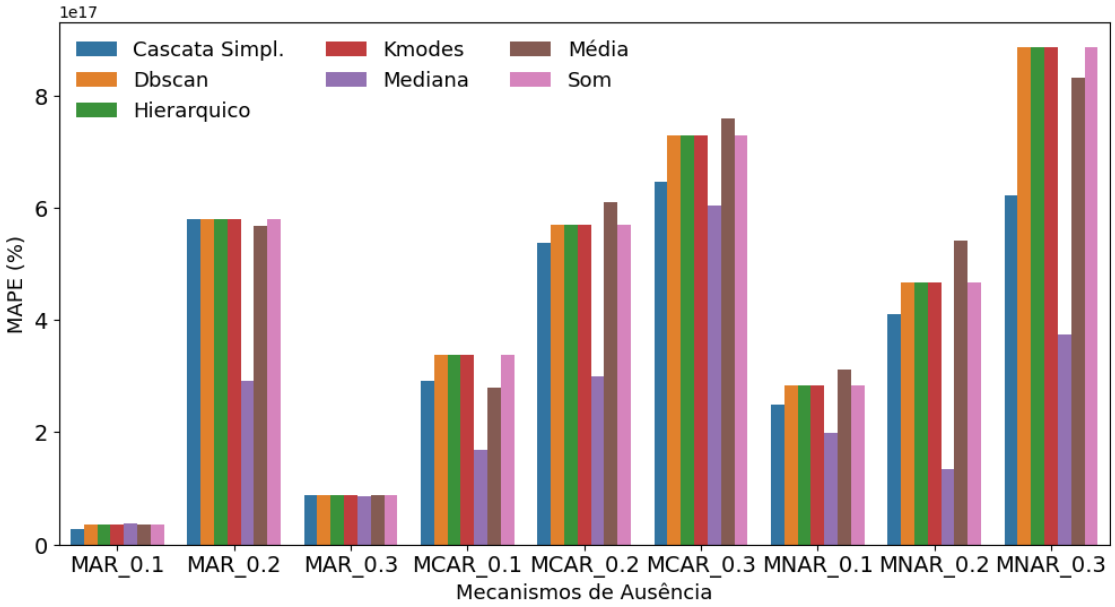


Figura V.28: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Pima*

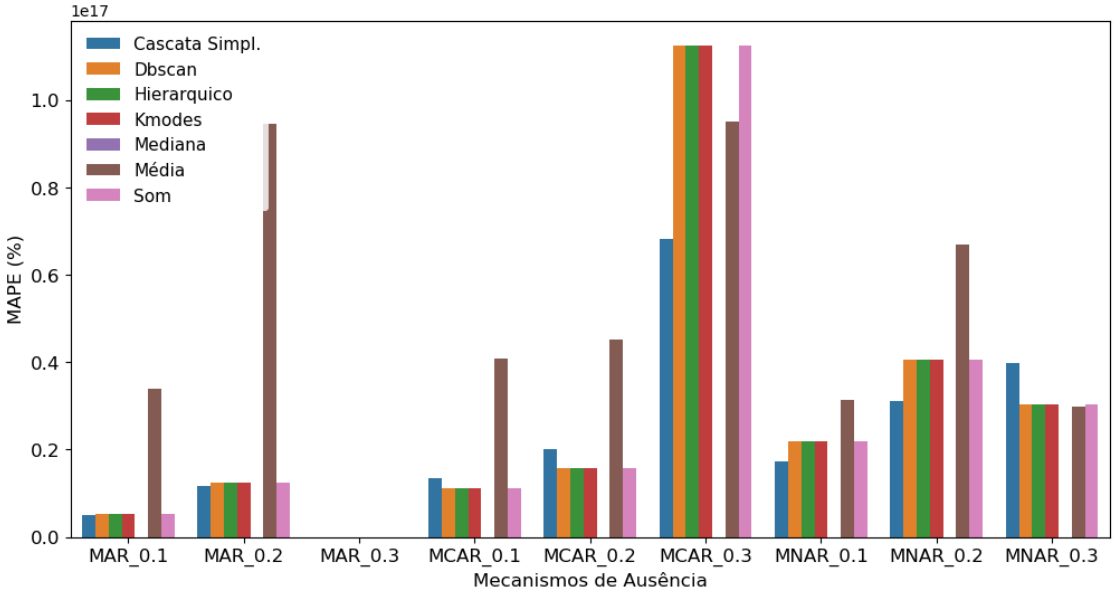


Figura V.29: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Boston*

Ao comparar os erros médios e mínimos de imputação da base *Boston*, nota-se que a diferença entre esses indicadores é pouco significativa. Contudo, ao analisar o erro médio (Figura V.29) nos mecanismos de ausência MCAR 20% e MNAR 30%, observa-se um desempenho inferior do método de Imputação em Cascata Simplificada. Em contrapartida, ao avaliarmos o erro mínimo (Figura V.30), esse método se destacou como o de melhor desempenho, apresentando o menor erro mínimo. No caso do mecanismo MCAR 30%, os métodos DBSCAN, Hierárquico, K-modes e Som apresentaram desempenho inferior ao da imputação pela média, conforme observado na análise dos erros médios (Figura V.29). Apesar disso, esses métodos demonstraram um desempenho superior nos mecanismos MCAR 10%, MCAR 20% e MNAR 30%. Por fim, constatou-se que a imputação por mediana foi a que apresentou o menor erro de imputação em todos os mecanismos de ausência na base *Boston*.

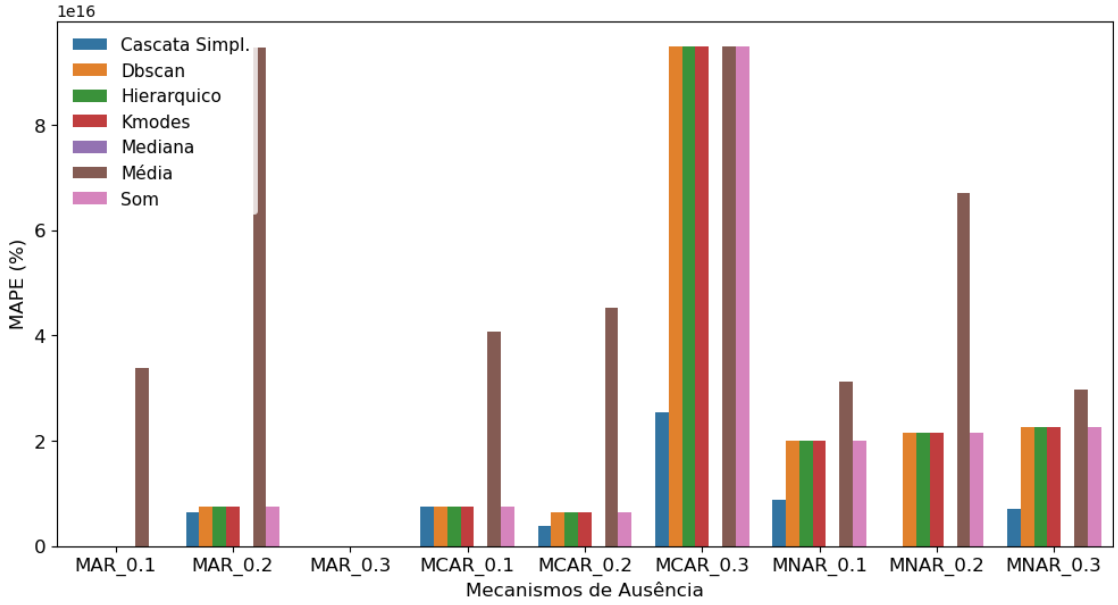


Figura V.30: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Boston*

Ao analisar os erros médios e mínimos de imputação da base *Abalone*, verifica-se que a diferença entre eles é pouco significativa, indicando que a distribuição dos erros é semelhante, conforme demonstrado nas Figuras V.31 e V.32. Todos os métodos apresentaram bom desempenho nos mecanismos de ausência MAR 20%, MCAR 10%, MCAR 20% e MNAR 20%. No mecanismo de ausência MAR 10%, o método de Imputação em Cascata Simplificada destacou-se tanto na análise do erro médio (Figura V.31) quanto no erro mínimo (Figura V.32). No entanto, no caso do MCAR 30%, esse método apresentou melhor desempenho apenas na análise do erro mínimo. Nos demais mecanismos de ausência, não houve diferenças significativas entre os métodos de Imputação em Cascata Simplificada, DBSCAN, Hierárquico, K-modes e Som, sendo que a imputação pela média e pela mediana apresentou os piores desempenhos.

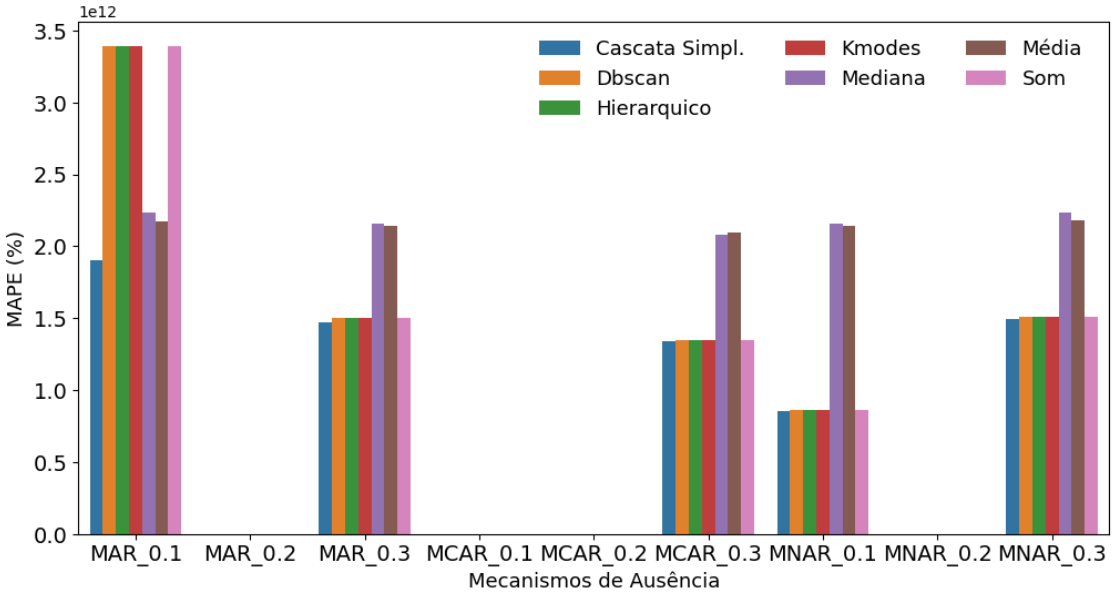


Figura V.31: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Abalone*

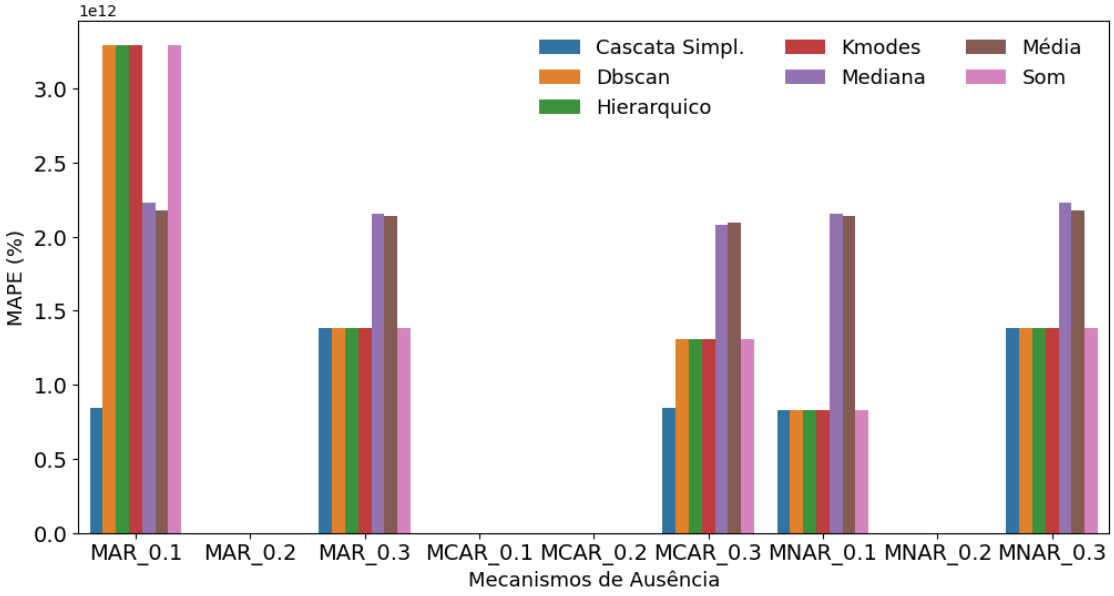


Figura V.32: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Abalone*

Ao comparar os erros médios e mínimos de imputação da base *Forest*, observa-se que, ao analisar o erro médio (Figura V.33) nos mecanismos de ausência MCAR 10%, MCAR 30% e MNAR 20%, o método de Imputação em Cascata Simplificada apresentou um desempenho inferior. No entanto, ao avaliarmos o erro mínimo (Figura V.34), esse método se destacou como o de melhor desempenho, apresentando o menor erro mínimo. É importante ressaltar que, na base *Forest*, todos os mecanismos de ausência MAR demonstraram um erro de imputação significativamente menor em comparação com os demais. De maneira geral, os métodos DBSCAN, Hierárquico, K-modes e Som apresentaram

desempenhos equivalentes.

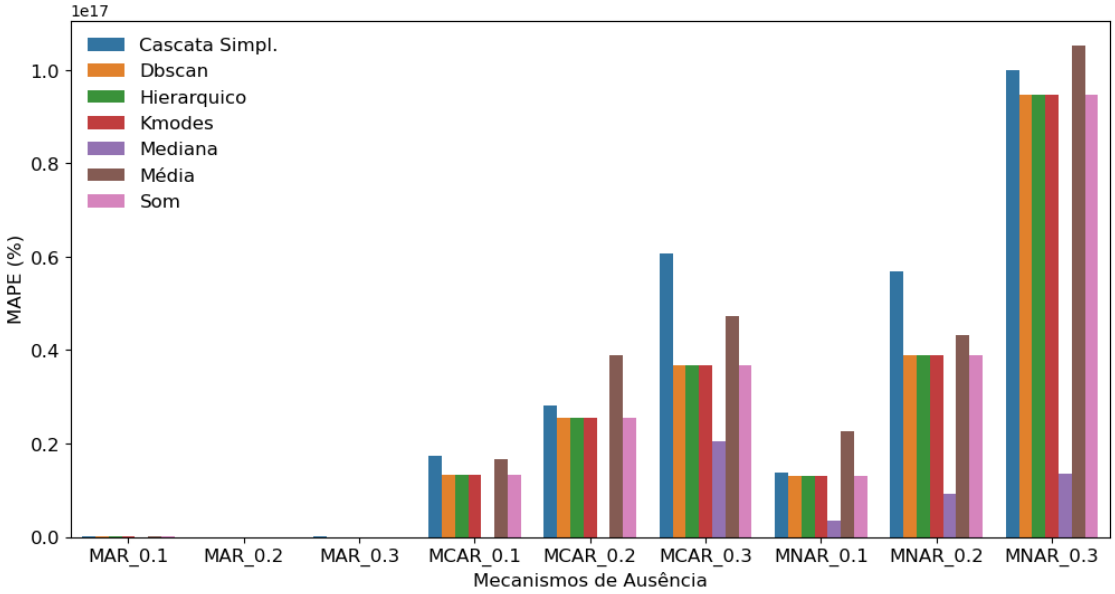


Figura V.33: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Forest*

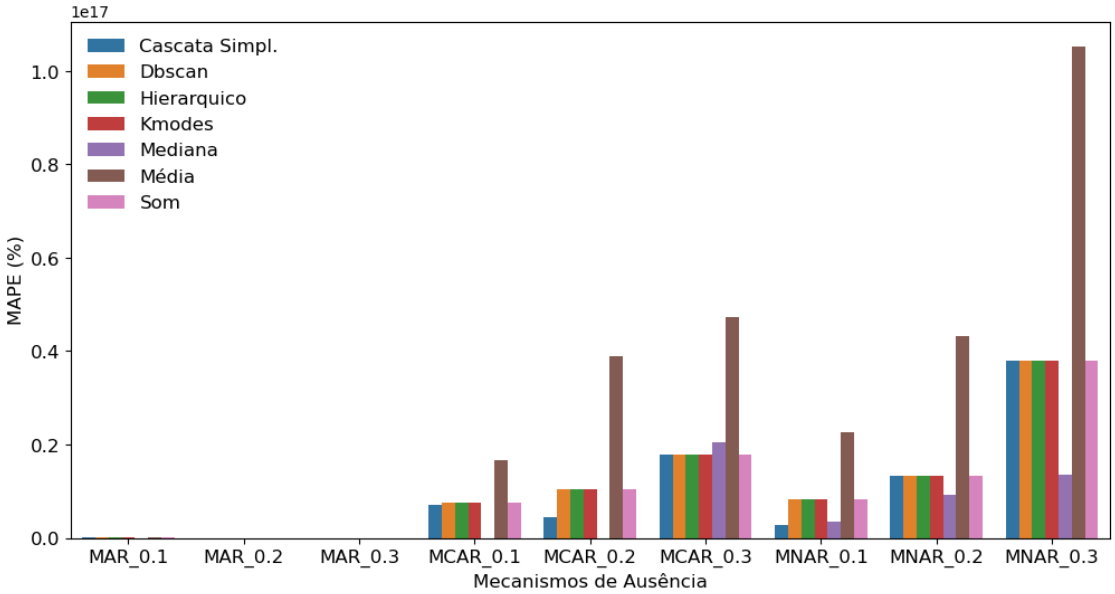


Figura V.34: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Forest*

Na base *Gamma*, tanto nos erros médios quanto nos mínimos, observa-se que, nos mecanismos de ausência MAR 30%, MCAR 10%, MCAR 20%, MNAR 20% e MNAR 30%, a imputação por mediana obteve um desempenho superior (Figura V.35 e V.36). Nos demais mecanismos de ausência, não houve diferenças significativas entre os erros dos métodos de imputação em Cascata Simplificada, DBSCAN, Hierárquico, K-modes e Som. Devido à similaridade no comportamento dos erros médios e mínimos, optou-se por apresentar apenas um gráfico, o qual sintetiza as informações de ambos os

tipos de erro.

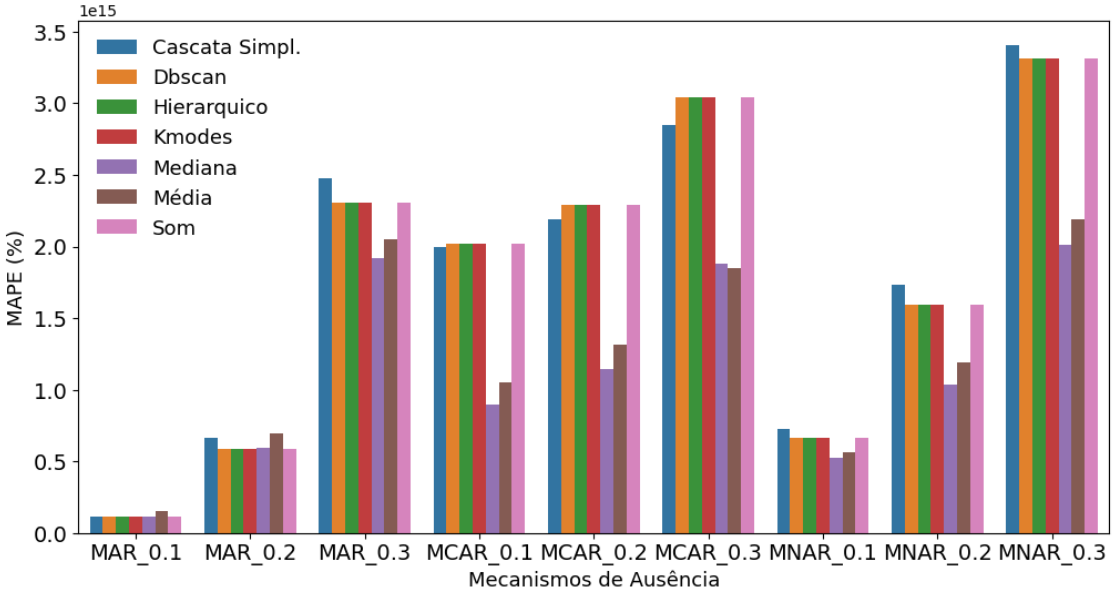


Figura V.35: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Gamma*

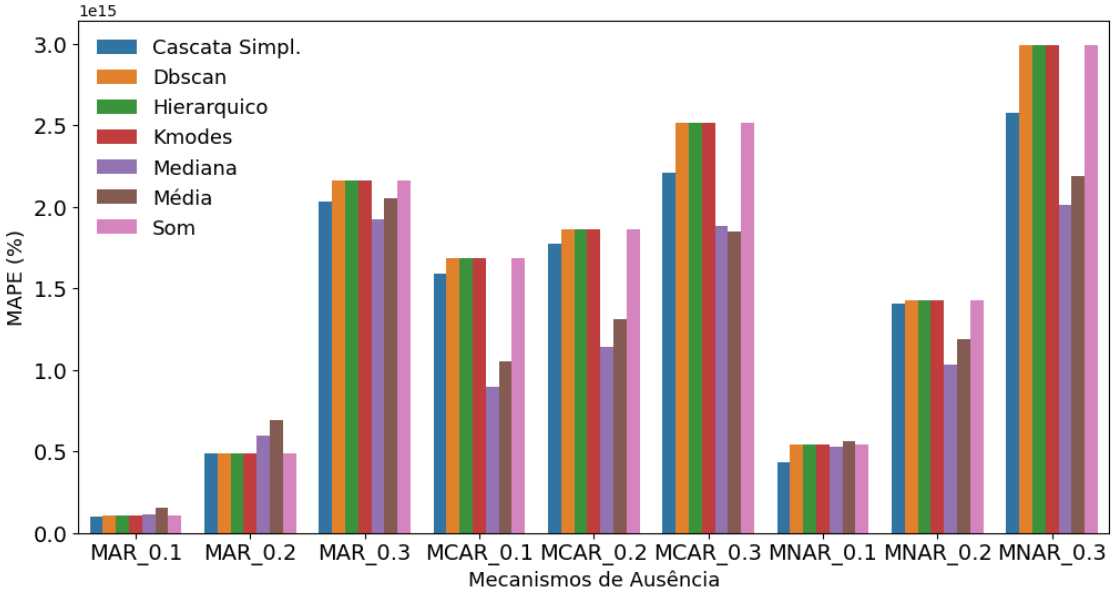


Figura V.36: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Gamma*

Ao analisar os erros médios (Figura V.37) e mínimos (Figura V.38) de imputação da base *Glass Identification*, verifica-se que a diferença entre esses indicadores é pouco significativa, indicando que a distribuição dos erros é semelhante. Além disso, exceto nos casos de MAR 20% e MNAR 10%, o método de Imputação em Cascata Simplificada apresentou o melhor desempenho. No mecanismo de ausência MAR 20%, a imputação por mediana apresentou melhores resultados, enquanto no caso do MNAR 10%, os métodos DBSCAN, Hierárquico, K-modes e SOM obtiveram os menores erros.

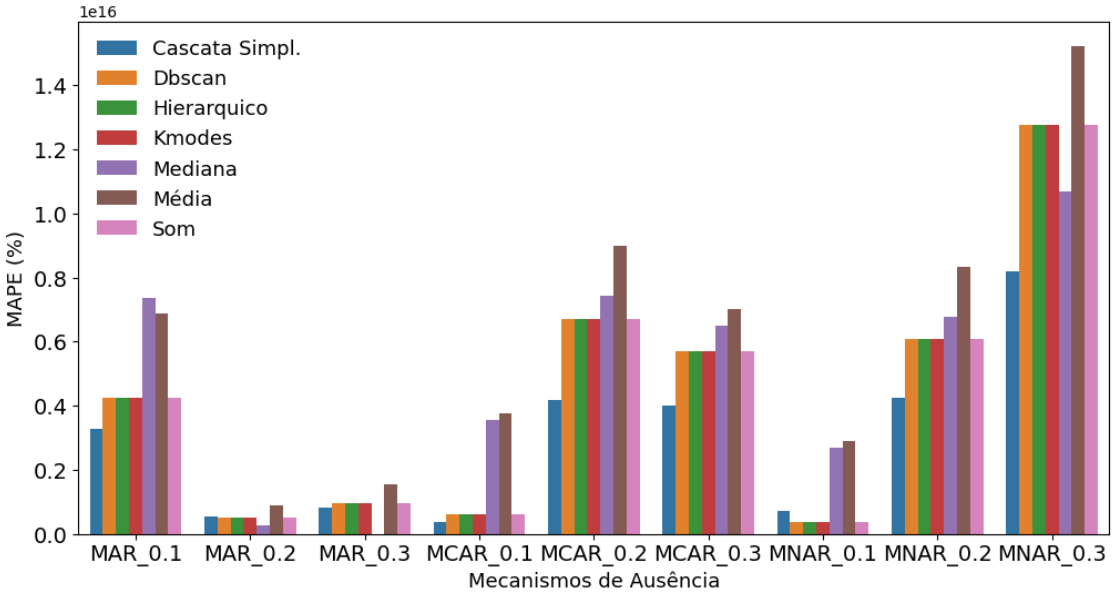


Figura V.37: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Glass*

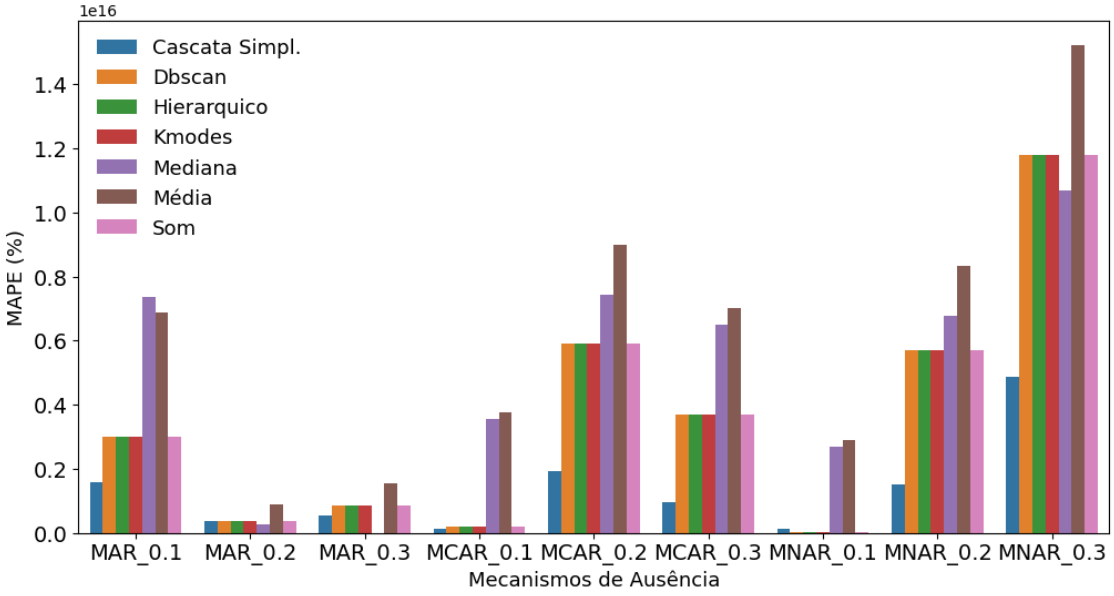


Figura V.38: Comparação dos erros míinimos de imputação (%) por método de imputação nos dados do *Glass*

Na base *Yeast*, a maior variação entre os erros médios (Figura V.39) e os erros mínimos (Figura V.40) ocorreu no método de imputação em Cascata Simplificada, especificamente no mecanismo de ausência MNAR 30%. Ao analisar o erro médio, observa-se que, nos mecanismos MCAR 10%, MCAR 20%, MNAR 10% e MNAR 30%, os métodos DBSCAN, Hierárquico, K-modes e Som apresentaram melhor desempenho. No entanto, ao considerar os erros mínimos, a imputação por mediana revelou-se a mais eficaz, exceto no caso do MNAR 30%.

Sumarizando as conclusões desta seção, a análise dos erros médios e mínimos de imputação

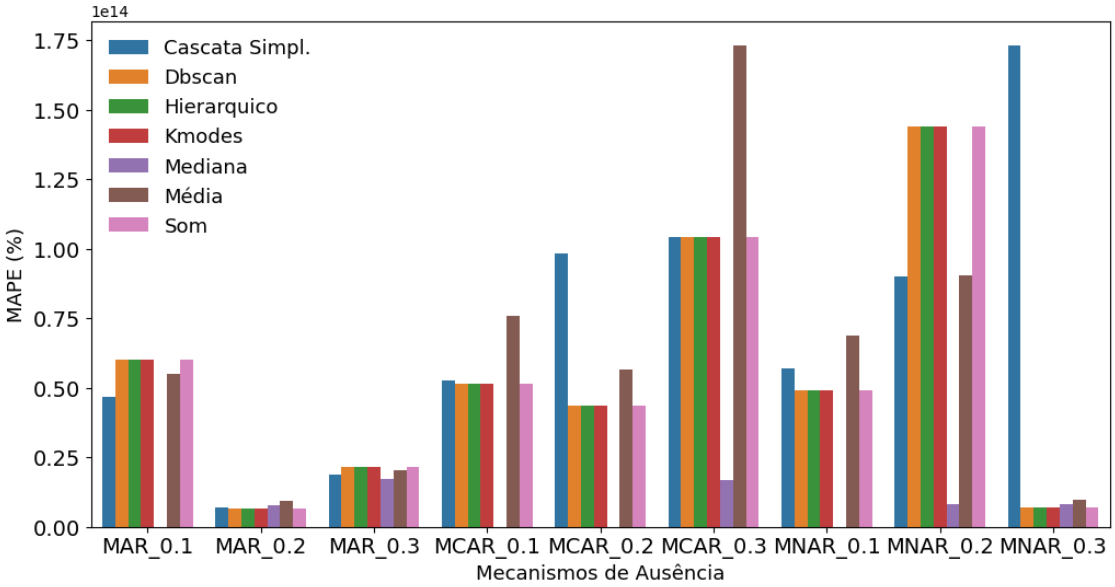


Figura V.39: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Yeast*

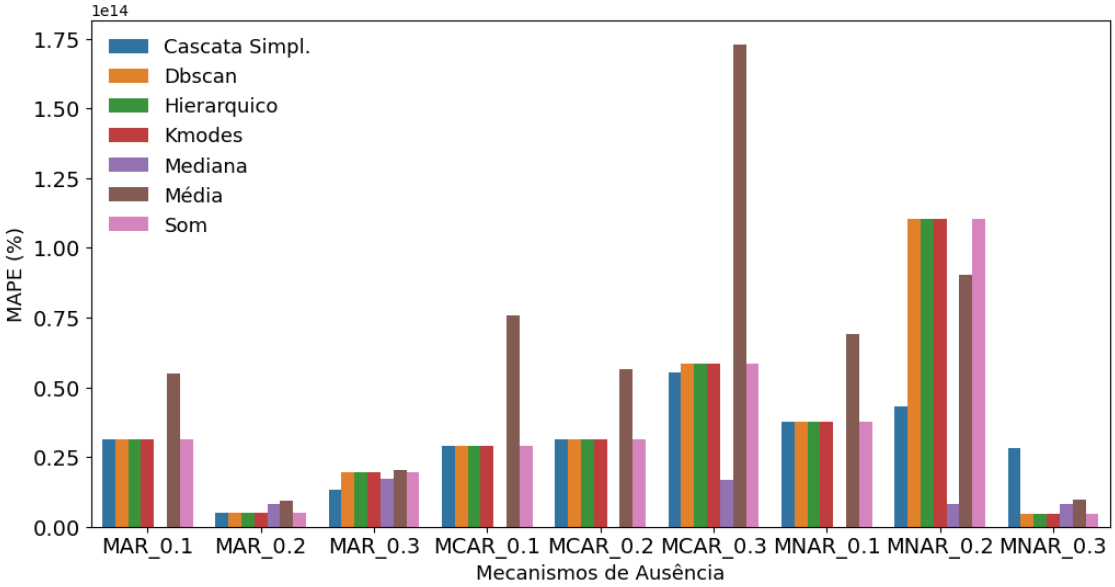


Figura V.40: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Yeast*

nas oito bases de dados forneceu insights valiosos sobre o desempenho dos diferentes métodos de imputação. De maneira geral, observou-se que o método de Imputação em Cascata Simplificada se destacou em diversas bases, apresentando o menor erro mínimo, especialmente nas bases *Forest* e *Boston*. Em contrapartida, a imputação por mediana demonstrou desempenho superior em bases como *Gamma*, particularmente nos mecanismos de ausência com maior percentual de dados faltantes, como MCAR 20% e MNAR 30%. Já os métodos DBSCAN, Hierárquico, K-modes e Som apresentaram desempenhos variados, com algumas bases mostrando resultados comparáveis, ou até superiores, à imputação pela média, dependendo do mecanismo de ausência utilizado.

Com base nessa análise, é possível concluir que, apesar das variações observadas entre os métodos, a imputação por mediana se mostrou uma abordagem robusta em diversos cenários, especialmente quando o erro mínimo foi o critério predominante de avaliação. No geral, os mecanismos de ausência MCAR e MNAR tiveram as diferenças mais significativas do desempenho entre os métodos de imputação, sugerindo que a escolha do método pode ser altamente sensível ao tipo e à proporção das ausências nos dados.

Avaliação do Bias de Correlação na Imputação em Cascata

Os gráficos desta seção comparam o impacto dos métodos de imputação testados na correlação original dos atributos, utilizando como métrica o índice denominado Bias da Correlação, conforme definido na seção II.4.3. A alteração da correlação original dos dados é uma consequência possível da reutilização dos valores durante o processo de imputação em Cascata. O ideal para essa métrica é que ela se aproxime de zero, indicando que não houve alteração significativa na correlação original.

Ao analisar o Bias de Correlação da base *Wine*, observa-se que a distribuição dos resultados médios, ou seja, o bias médio (Figura V.41), e dos resultados mínimos, ou seja, o bias mínimo (Figura V.42), mantém-se consistente, apresentando as mesmas características nos dois casos. Além disso, foi possível identificar que os mecanismos de ausência MCAR 20%, MCAR 30% e MNAR 30% geraram as maiores correlações. Em outras palavras, nesses mecanismos de ausência, todos os métodos de imputação alteraram a correlação original dos dados em pelo menos 18%. A imputação pela média e pela mediana também resultaram em um bias de correlação elevado.



Figura V.41: Comparação do bias de correlação médio entre os dados originais de *Wine* e os dados imputados



Figura V.42: Comparação do bias de correlação mínimo entre os dados originais de *Wine* e os dados imputados

Na base *Pima*, observamos que todos os mecanismos de ausência apresentaram um bias de correlação superior ao registrado na base *Wine*, tanto para o bias médio (Figura V.43) quanto para o bias mínimo (Figura V.44). Além disso, verificou-se que a imputação pela média e pela mediana geraram os menores bias de correlação para este conjunto de dados.

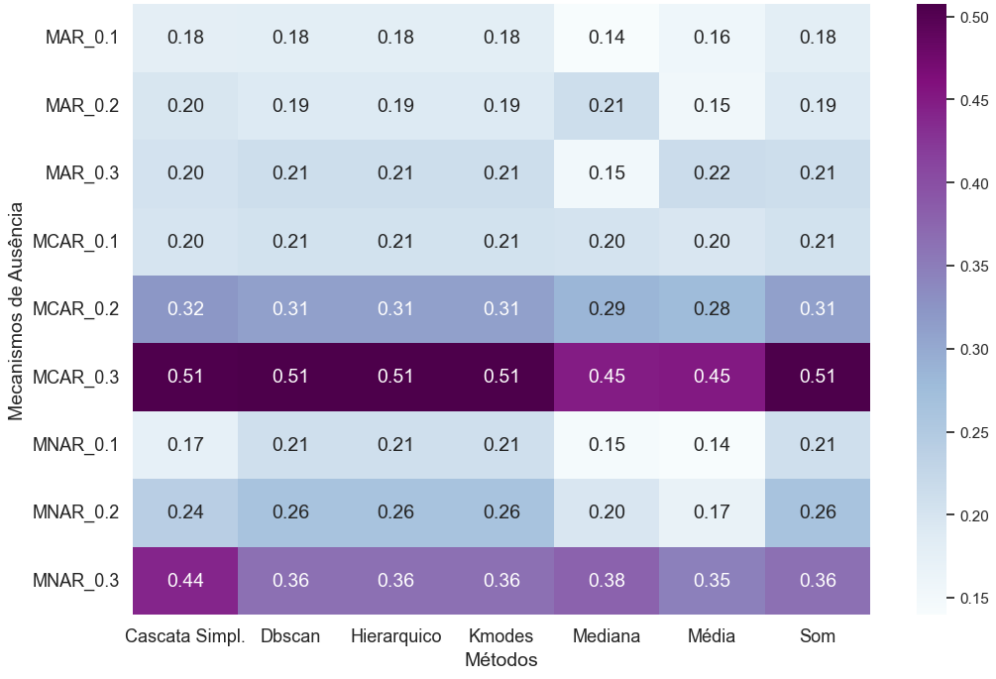


Figura V.43: Comparação do bias de correlação médio entre os dados originais de *Pima* e os dados imputados

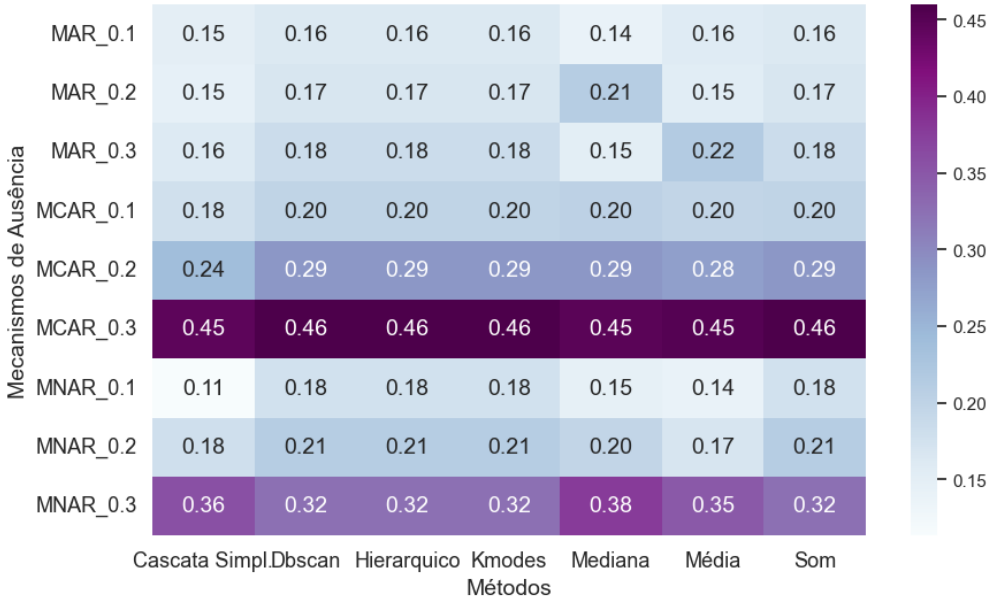


Figura V.44: Comparação do bias de correlação mínimo entre os dados originais de *Pima* e os dados imputados

Ao analisar o Bias de Correlação da base *Boston*, a distribuição dos resultados médios, ou seja, o bias médio (Figura V.45), e dos resultados mínimos, ou seja, o bias mínimo (Figura V.46), apresentou características semelhantes entre os resultados. Observou-se também que, com exceção da imputação pela média e pela mediana no MAR 20%, que apresentaram bias de correlação de -37% e -34%, respectivamente, os demais métodos geraram valores baixos de bias, indicando que a imputação não alterou significativamente a correlação original do dataset.

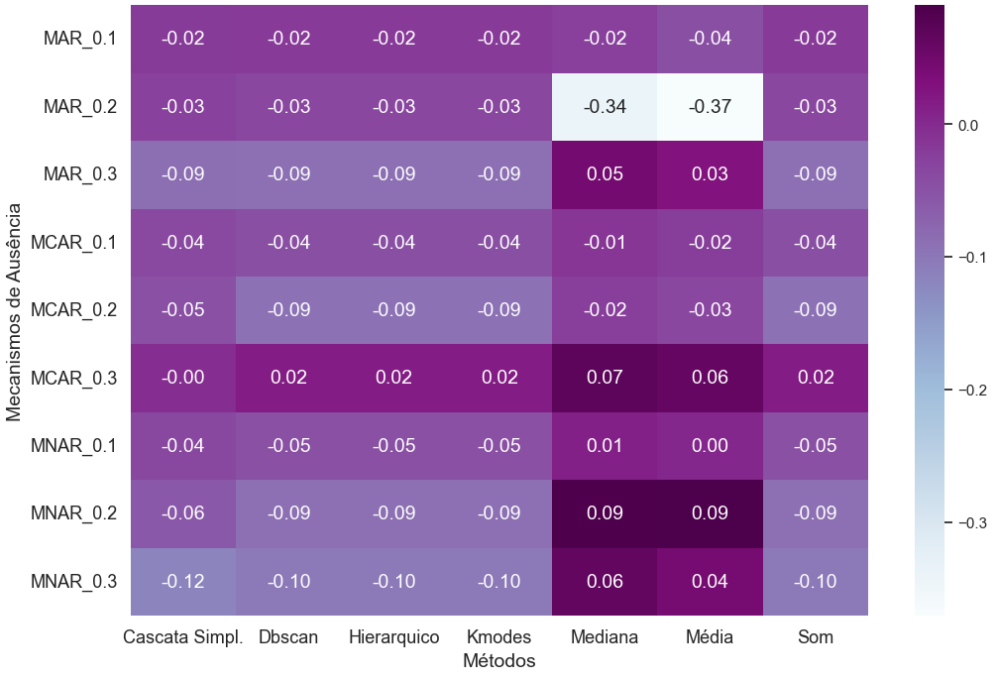


Figura V.45: Comparação do bias de correlação médio entre os dados originais de *Boston* e os dados imputados

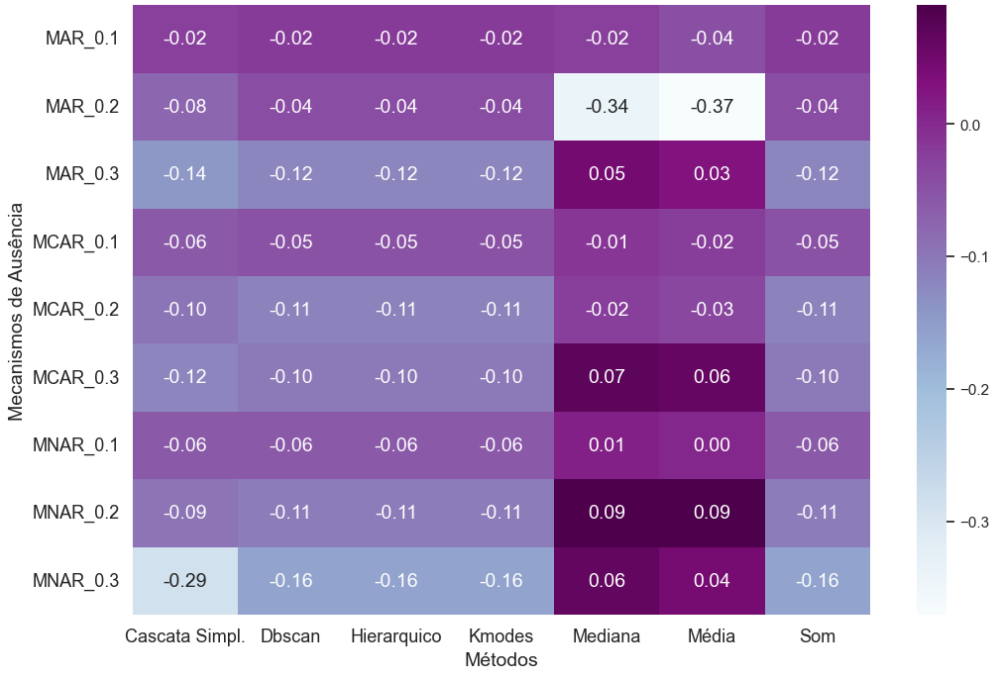


Figura V.46: Comparação do bias de correlação mínimo entre os dados originais de *Boston* e os dados imputados

Na base *Abalone*, as características e a distribuição dos resultados para o bias médio (Figura V.47) e o bias mínimo (Figura V.48) também se mantiveram consistentes, com os resultados apresentando as mesmas características. Para este dataset, os melhores resultados foram observados nos mecanismos de ausência MAR 10%, MCAR 10% e MNAR 10%, em todos os métodos testados, com exceção das imputações univariadas (média e mediana).

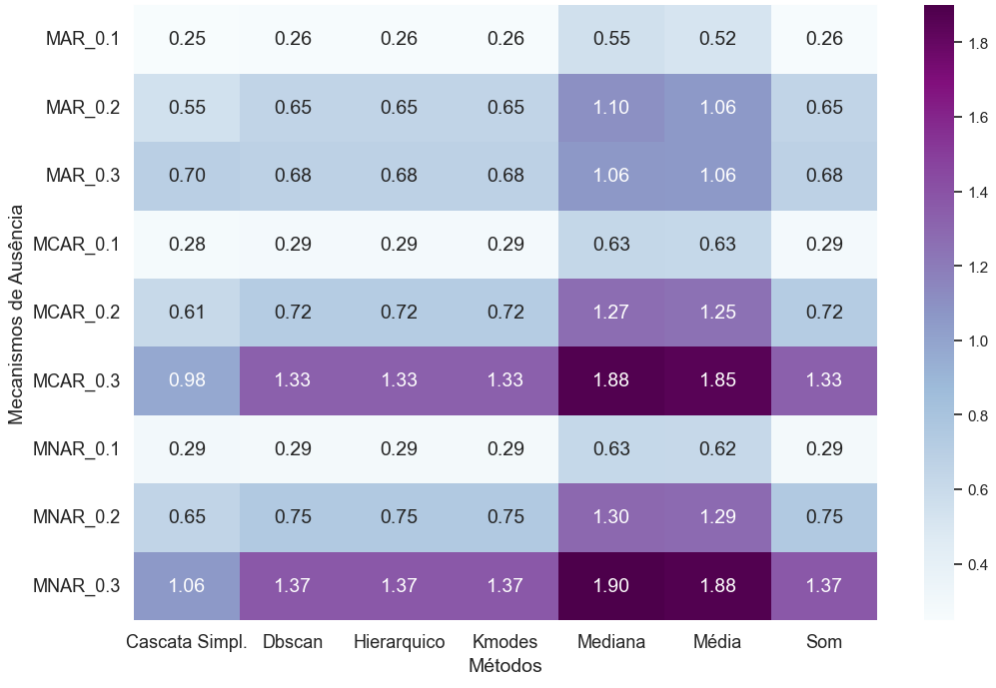


Figura V.47: Comparação do bias de correlação médio entre os dados originais de *Abalone* e os dados imputados



Figura V.48: Comparação do bias de correlação mínimo entre os dados originais de *Abalone* e os dados imputados

Na base *Forest*, observou-se que a distribuição do bias médio (Figura V.49) e do bias mínimo (Figura V.50) permaneceu consistente entre os resultados. As imputações por média e mediana geraram os maiores valores de bias médio e bias mínimo. Além disso, foi possível identificar que os melhores resultados foram obtidos nos mecanismos de ausência MAR 10%, MAR 20%, MCAR 10%, com exceção das imputações univariadas (média e mediana).



Figura V.49: Comparação do bias de correlação médio entre os dados originais de *Forest* e os dados imputados

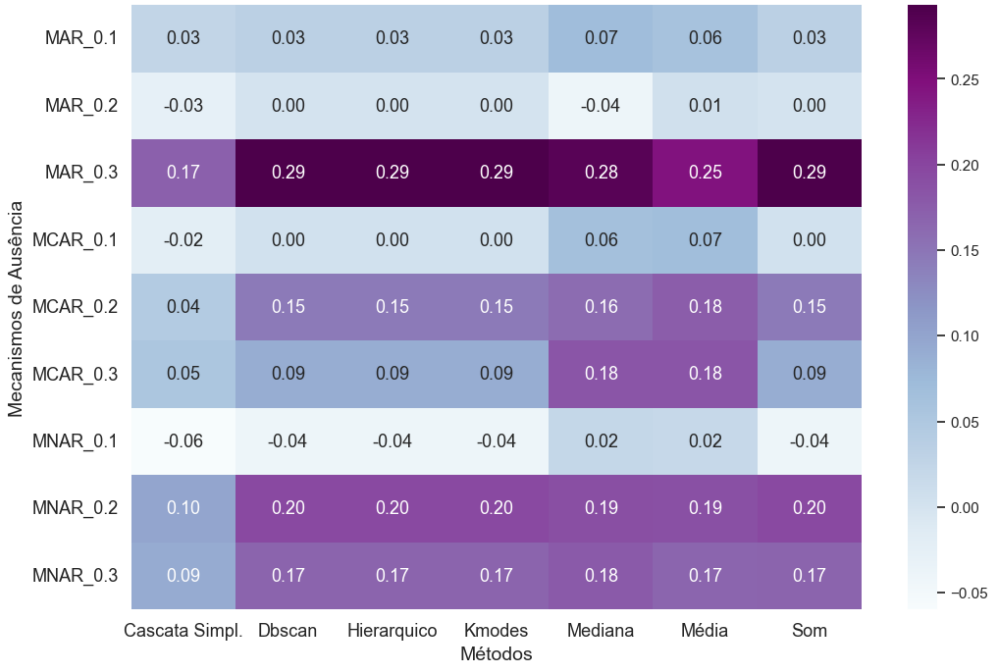


Figura V.50: Comparação do bias de correlação mínimo entre os dados originais de *Forest* e os dados imputados

Na base *Gamma*, verificou-se uma diferença entre os resultados do bias médio (Figura V.51) e do bias mínimo (Figura V.52), que variaram entre -4% e 0%. Para esse conjunto de dados, os melhores resultados foram obtidos nos mecanismos de ausência MAR 10%, MCAR 10% e MNAR 10%, independentemente do método de imputação utilizado. No entanto, ao comparar os métodos, observou-se que a imputação pela média e pela mediana apresentaram o melhor desempenho.

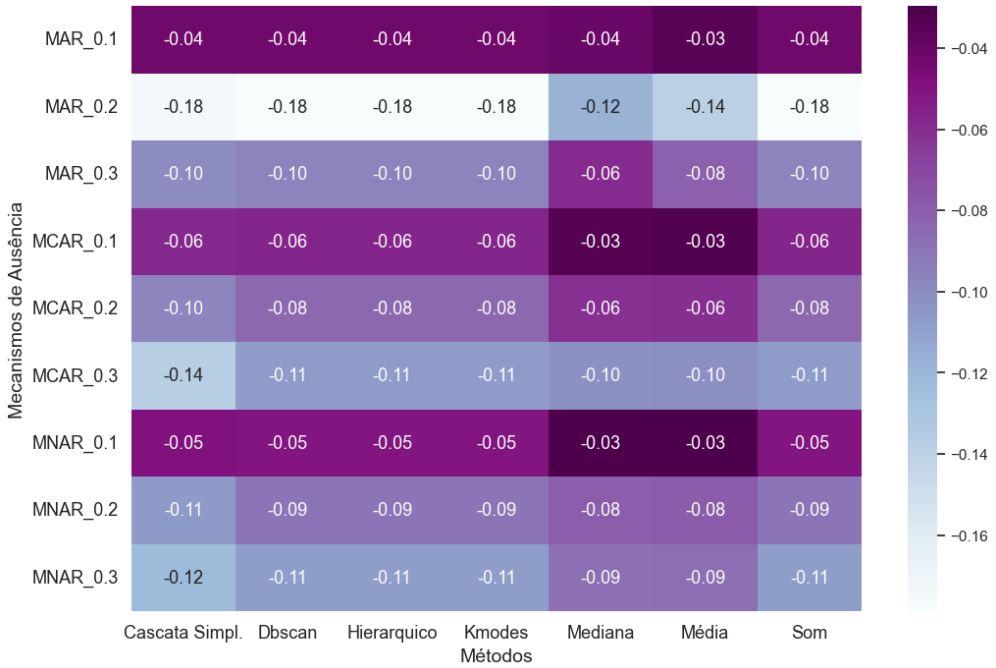


Figura V.51: Comparação do bias de correlação médio entre os dados originais de *Gamma* e os dados imputados

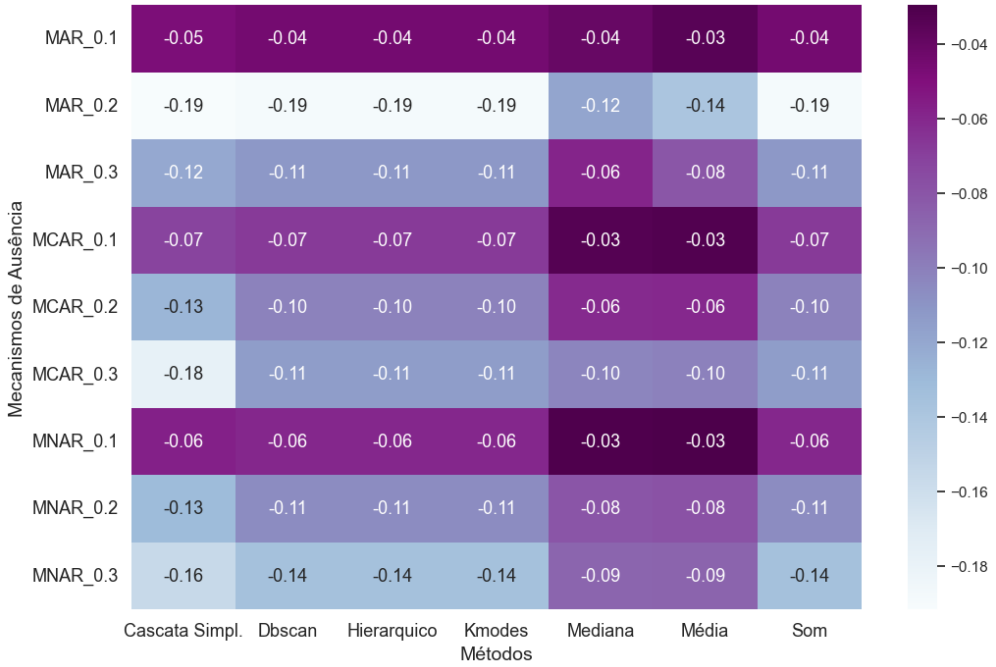


Figura V.52: Comparação do bias de correlação mínimo entre os dados originais de *Gamma* e os dados imputados

Ao analisar os resultados do bias de correlação na base *Glass*, constatou-se que os mecanismos de ausência MAR 10%, MAR 20%, MCAR 10% e MNAR 10% apresentaram bias de correlação mais próximos de zero, tanto no bias médio (Figura V.37) quanto no bias mínimo (Figura V.38). Nesse caso, os métodos de imputação pela média e pela mediana apresentaram os piores resultados, enquanto a imputação em Cascata Simplificada obteve o melhor desempenho.

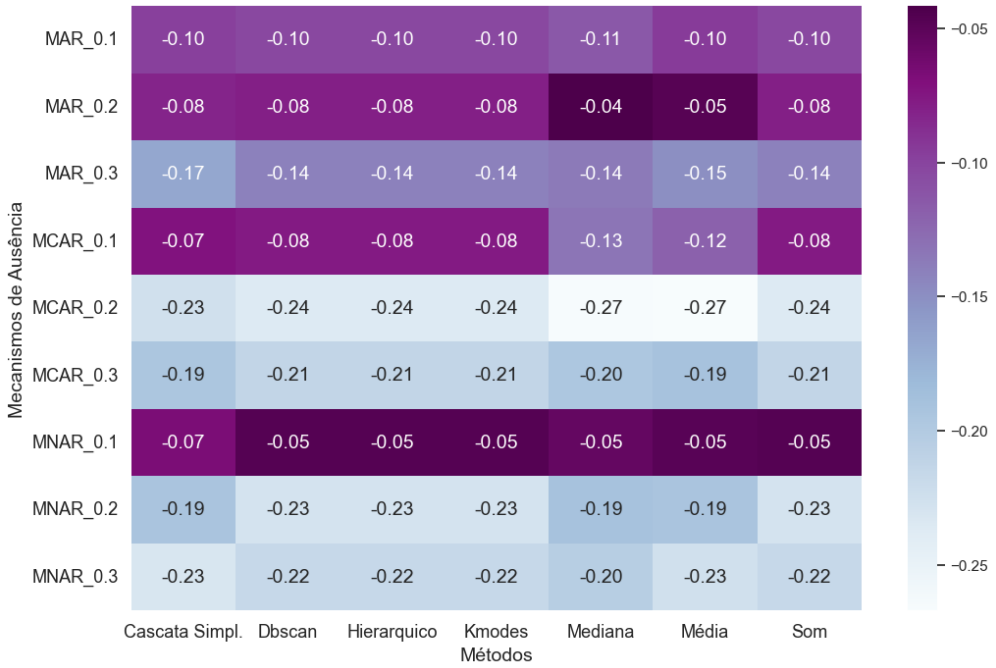


Figura V.53: Comparação do bias de correlação médio entre os dados originais de *Glass* e os dados imputados

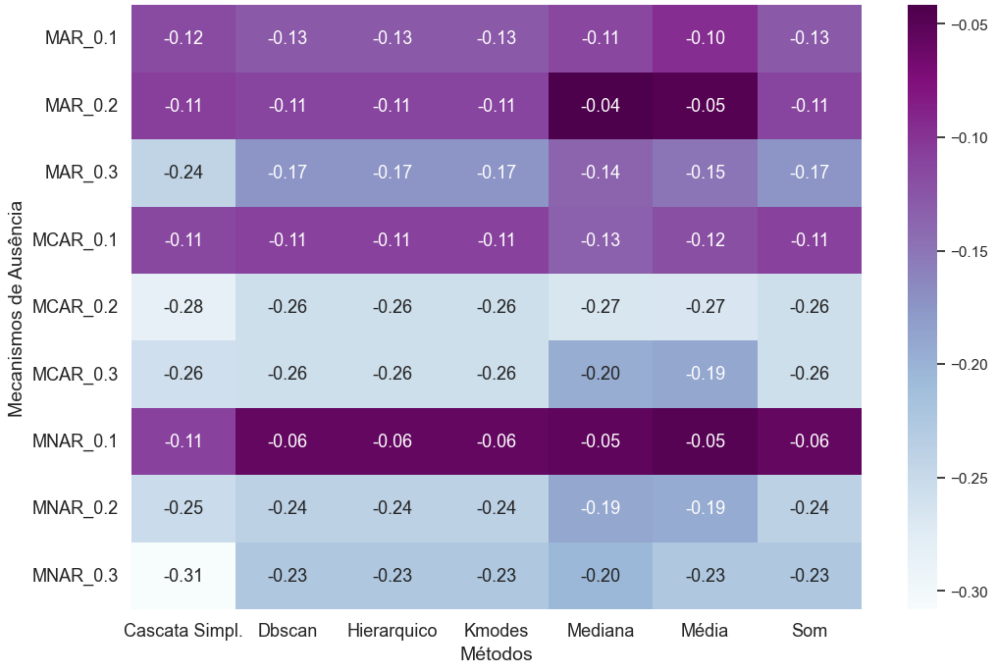


Figura V.54: Comparação do bias de correlação mínimo entre os dados originais de *Glass* e os dados imputados

Por fim, na base *Yeast*, observou-se que não há diferença significativa entre o bias de correlação médio (Figura V.55) e o bias de correlação mínimo (Figura V.56). O bias de correlação apresentou valores consistentemente baixos, sempre inferiores a 10%. Para este conjunto de dados, o desempenho foi insatisfatório nos mecanismos de ausência MAR 30% e MCAR 30%. No entanto, todos os métodos testados mostraram bons resultados nos demais cenários de mecanismos de ausência.

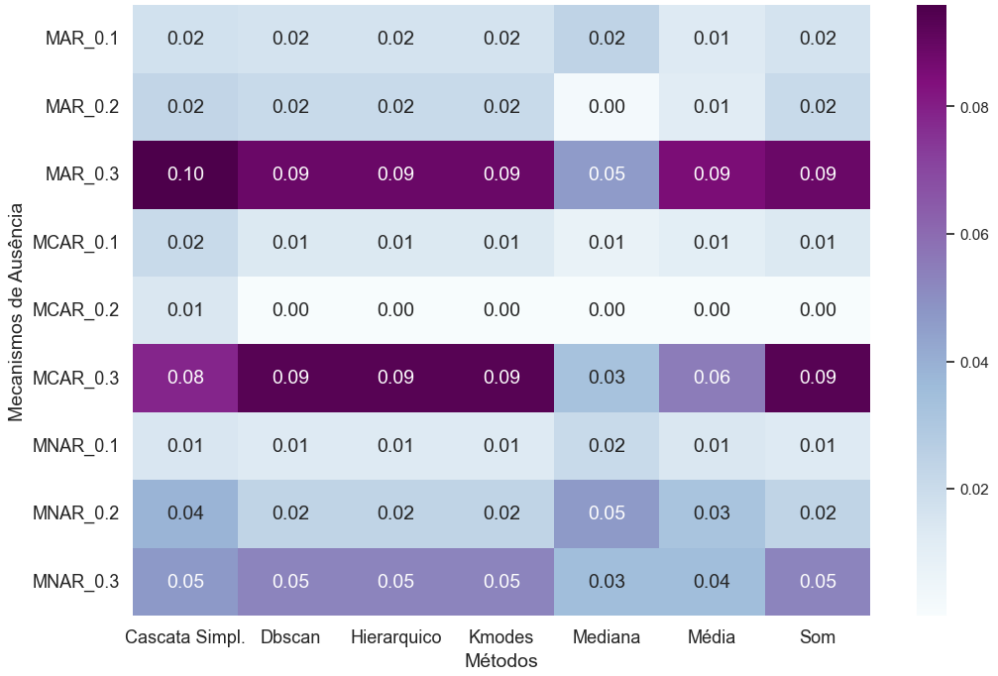


Figura V.55: Comparação do bias de correlação médio entre os dados originais de *Yeast* e os dados imputados

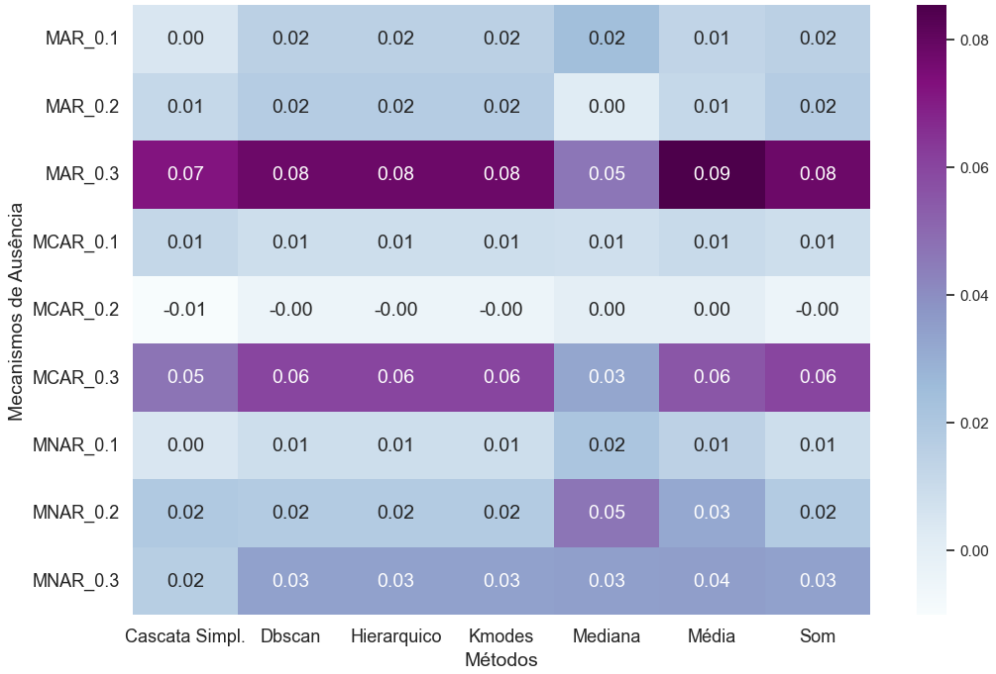


Figura V.56: Comparação do bias de correlação mínimo entre os dados originais de *Yeast* e os dados imputados

Em conclusão, a análise dos bias de correlação nas diferentes bases de dados revelou que, de maneira geral, a imputação em Cascata Simplificada apresentou um desempenho consistente, ou seja, os valores de bias médio e mínimo próximos de zero, indicando pouca alteração na correlação original dos dados. As bases *Forest*, *Gamma* e *Glass* mostraram que os mecanismos de ausência, como MAR 10%, MCAR 10% e MNAR 10%, foram os que apresentaram os menores bias de correlação, independentemente do método de imputação utilizado. No entanto, para as bases *Wine* e *Pima*, a imputação por média e mediana teve um desempenho mais significativo em termos de alteração da correlação original, resultando em valores mais elevados de bias.

Além disso, a análise indicou que, em várias bases, a imputação univariada (média e mediana) frequentemente resultou nos maiores bias de correlação, o que sugere que esses métodos alteram substancialmente as relações entre os atributos. A escolha do método de imputação, portanto, deve considerar não apenas a precisão na substituição dos valores ausentes, mas também seu impacto na estrutura das correlações entre os atributos. A Imputação em Cascata, por sua vez, demonstrou ser uma abordagem eficaz para preservar as correlações originais, especialmente em datasets com uma distribuição de ausências mais equilibrada.

Avaliação do Erro de Similaridade na Imputação em Cascata

Os gráficos desta seção comparam o impacto dos métodos de imputação testados na eficácia da imputação. Através dessa análise, será possível avaliar o quão próximos os valores imputados estão dos valores verdadeiros. O ideal para essa métrica é que ela se aproxime de zero, o que indicaria

que os dados permanecem semelhantes após a imputação. Embora nas seções anteriores tenhamos comparado os valores médios e mínimos, nesta seção optamos por não realizá-la, uma vez que em todas as bases não foram observadas diferenças significativas entre esses dois indicadores.

Ao analisar o erro de similaridade na base *Wine*, observou-se que, em todos os mecanismos de ausência, o método de imputação em Cascata Simplificada obteve os menores erros (Figura V.57). Além disso, foi identificado que, dentro dos mecanismos MAR, MCAR e MNAR, à medida que a proporção de valores ausentes aumentava, o erro de similaridade também aumentava, o que é consistente, considerando que uma maior quantidade de dados ausentes resulta em uma redução da base original de comparação.

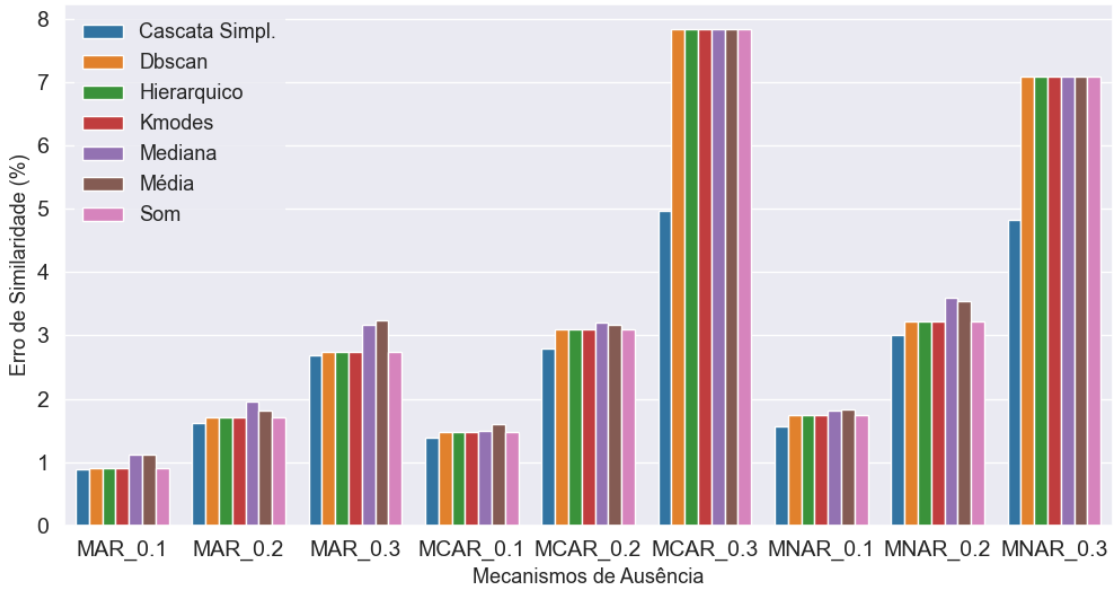


Figura V.57: Comparação da similaridade média entre a base original e a imputada na base *Wine*

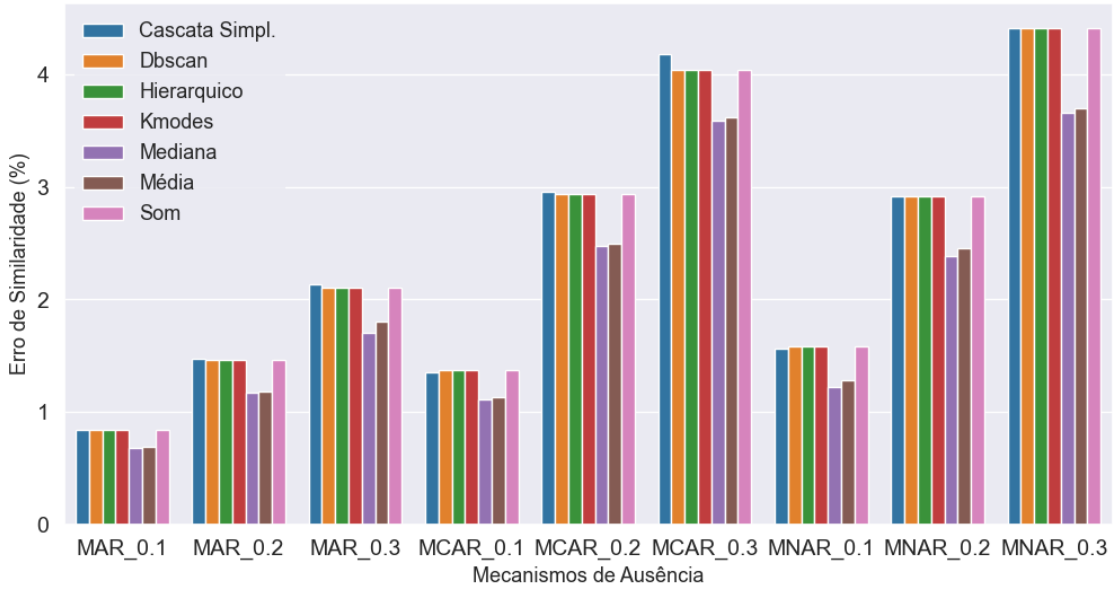


Figura V.58: Comparação da similaridade média entre a base original e a imputada na base *Pima*

Na base *Pima*, observou-se que o comportamento dos erros foi semelhante em todos os mecanismos de ausência (Figura V.58). Dessa forma, pode-se afirmar que os métodos de imputação por média e mediana foram os que melhor mantiveram a similaridade com a base original.

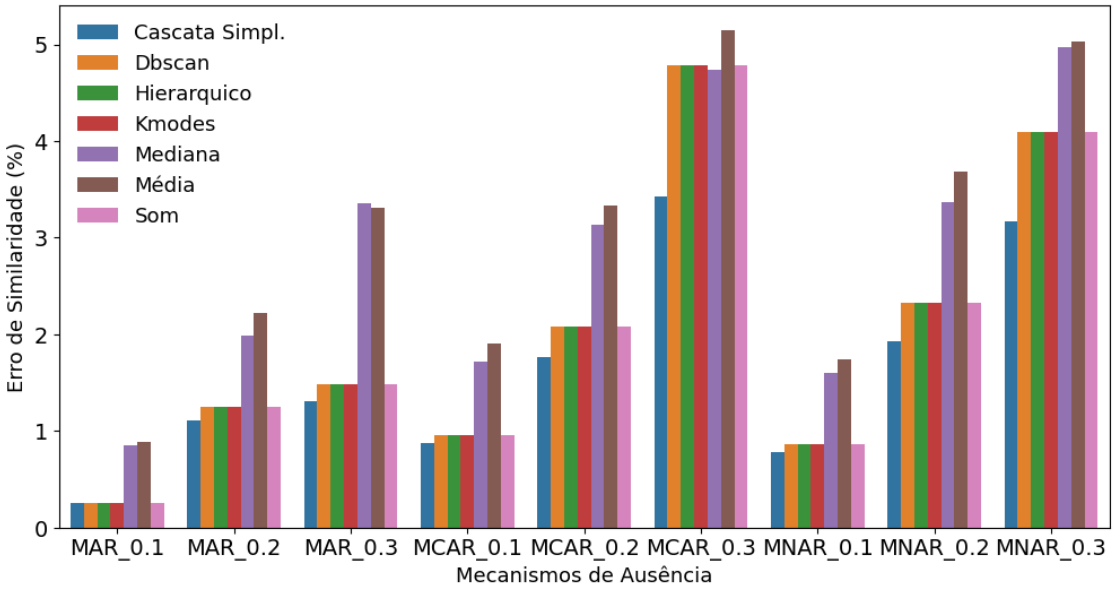


Figura V.59: Comparação da similaridade média entre a base original e a imputada na base *Boston*

Na base *Boston*, com exceção do mecanismo MAR 10%, a Imputação em Cascata Simplificada foi o método que conseguiu manter a maior similaridade entre a base original e a imputada (Figura V.59).

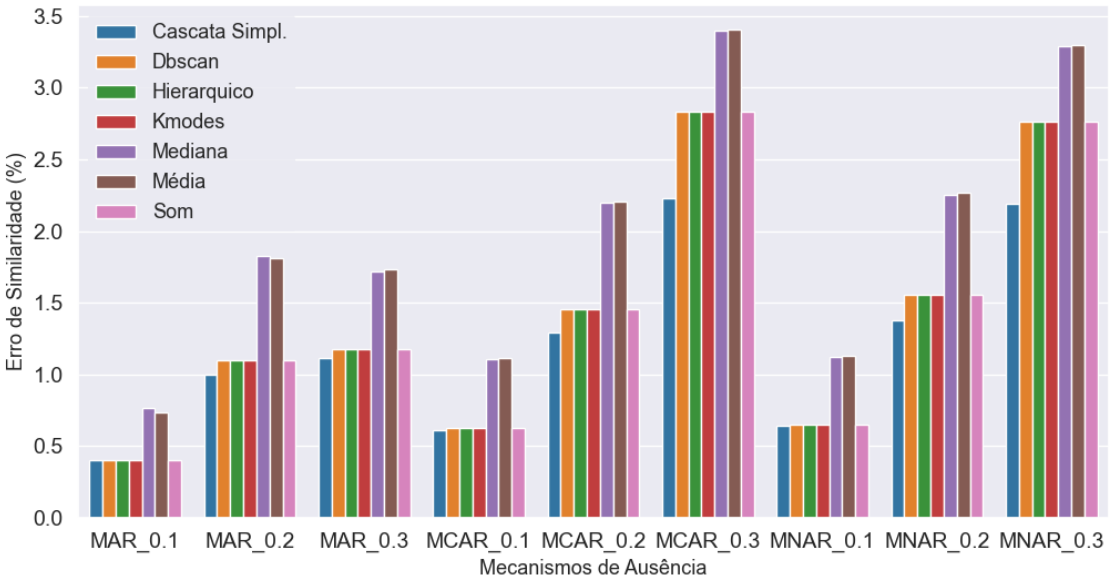


Figura V.60: Comparação da similaridade média entre a base original e a imputada na base *Abalone*

Em relação à base *Abalone*, o método de Imputação em Cascata Simplificada obteve os melhores resultados, exceto nos mecanismos de ausência MAR 10%, MCAR 10% e MNAR 10%. Em contrapartida, os métodos de imputação univariada, como a média e a mediana, apresentaram o pior

desempenho (Figura V.60).

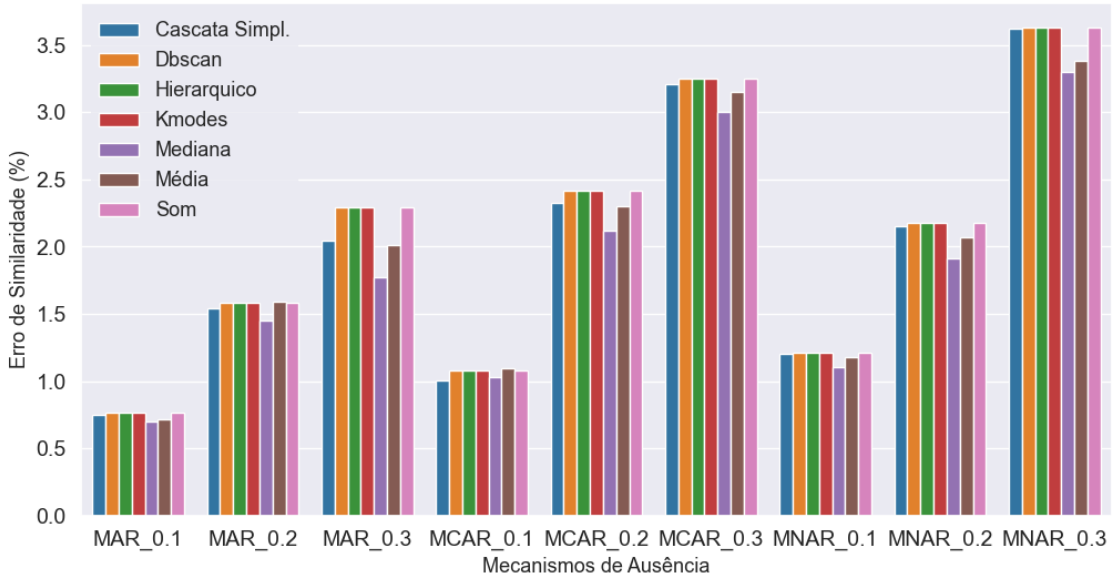


Figura V.61: Comparação da similaridade média entre a base original e a imputada na base *Forest*

Ao observar os dados da base *Forest* (Figura V.61), constatou-se que o método de Imputação em Cascata Simplificada obteve os melhores resultados no mecanismo MCAR 10%, enquanto a imputação por mediana foi o método com melhor desempenho nos demais mecanismos de ausência.

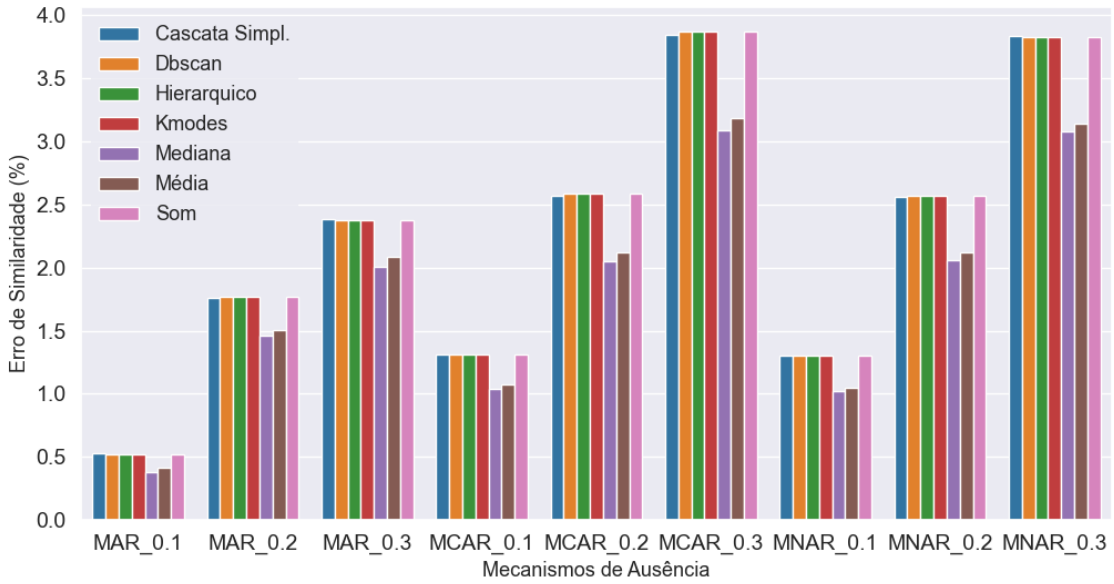


Figura V.62: Comparação da similaridade média entre a base original e a imputada na base *Gamma*

Na base *Gamma*, os métodos univariados (média e mediana) apresentaram o menor percentual de erro de similaridade, com a imputação por mediana sendo a mais eficaz das duas (Figura V.62). Os demais métodos não tiveram uma diferença significativa entre eles.

Quanto à base *Glass*, os métodos de imputação por média e mediana obtiveram os piores resultados em praticamente todos os mecanismos de ausência analisados (Figura V.63). Por outro

lado, o método de Imputação em Cascata Simplificada obteve o melhor desempenho em todos os mecanismos de ausência, exceto nos casos de MAR 20% e MNAR 10%.

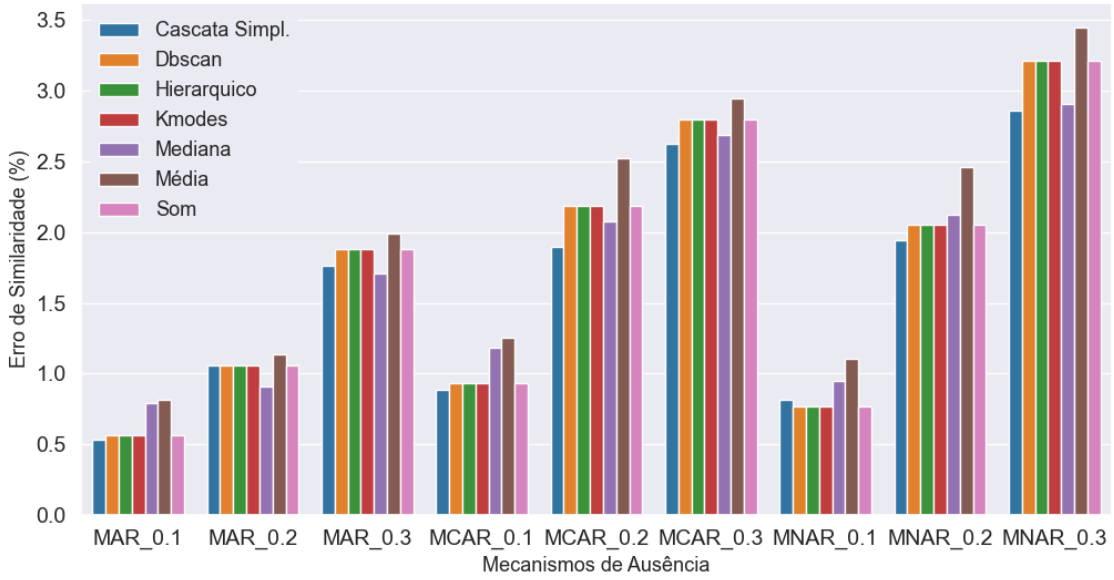


Figura V.63: Comparação da similaridade média entre a base original e a imputada na base *Glass*

Finalmente, na base *Yeast*, a imputação por mediana, seguida pela imputação por média, teve os melhores resultados em todos os mecanismos de ausência testados (Figura V.64). Em contrapartida, o método de Imputação em Cascata Simplificada apresentou o pior desempenho nos mecanismos MCAR 20% e MNAR 30%.

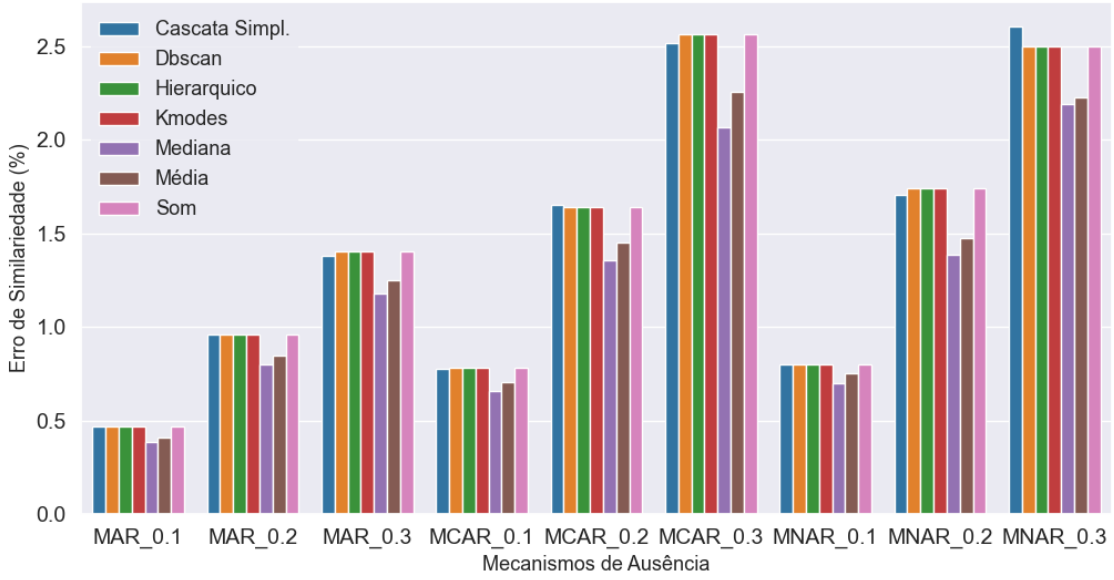


Figura V.64: Comparação da similaridade média entre a base original e a imputada na base *Yeast*

Em conclusão, a análise dos erros de similaridade revelou que o método de Imputação em Cascata Simplificada foi o mais eficaz em várias bases de dados, especialmente nos mecanismos de ausência MCAR 10% e MNAR 10%, onde obteve os melhores resultados de similaridade dos dados imputados

com a base original. No entanto, a imputação por média e mediana demonstrou um desempenho superior em determinadas bases, como *Pima* e *Gamma*.

A análise sugere que, em geral, a escolha do método de imputação tem um impacto direto na capacidade de manter a similaridade dos dados imputados com os dados originais, sendo que a Imputação em Cascata Simplificada se mostrou robusta na preservação dessa similaridade em muitos cenários. Contudo, em alguns casos, como na base *Yeast*, a imputação por mediana foi a mais eficaz, evidenciando que a combinação entre o tipo de dados e o mecanismo de ausência deve ser cuidadosamente considerada ao selecionar o método de imputação. Assim, a escolha do método adequado depende não apenas da precisão da imputação, mas também da capacidade de preservar as características fundamentais dos dados originais.

V.4.2 Comparação do desempenho dos Critério de Ordenação

Em duas fases distintas da Imputação em Cascata, é necessário definir a ordem de imputação. Primeiramente, os grupos devem ser ordenados, e, em seguida, deve-se determinar a ordem em que os atributos serão imputados. Existem oito possíveis ordens para os grupos e cinco ordens para os atributos, totalizando 40 configurações diferentes. A Tabela A.1, no apêndice A, resume essas configurações e fornece uma legenda para cada uma das combinações.

Nesta seção, compararemos o erro médio e o erro mínimo no desempenho dos critérios de ordenação. Os gráficos apresentam o critério de ordem no eixo x, organizado de forma crescente pelo percentual de erro, permitindo identificar as variações no desempenho dos métodos de imputação em relação à ordem dos grupos e atributos.

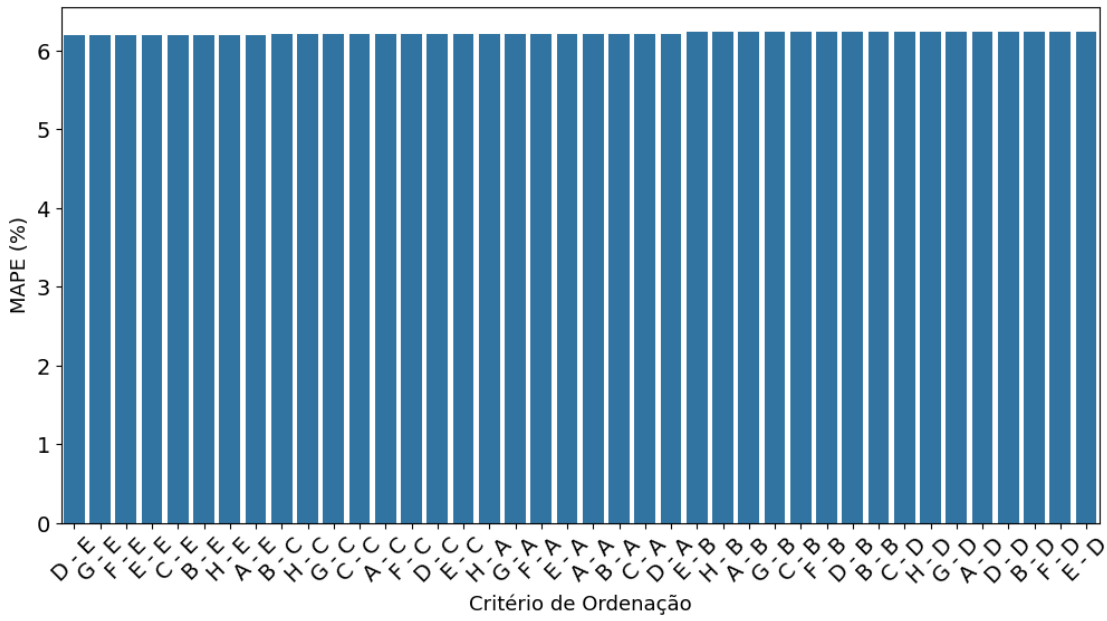


Figura V.65: Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Wine*

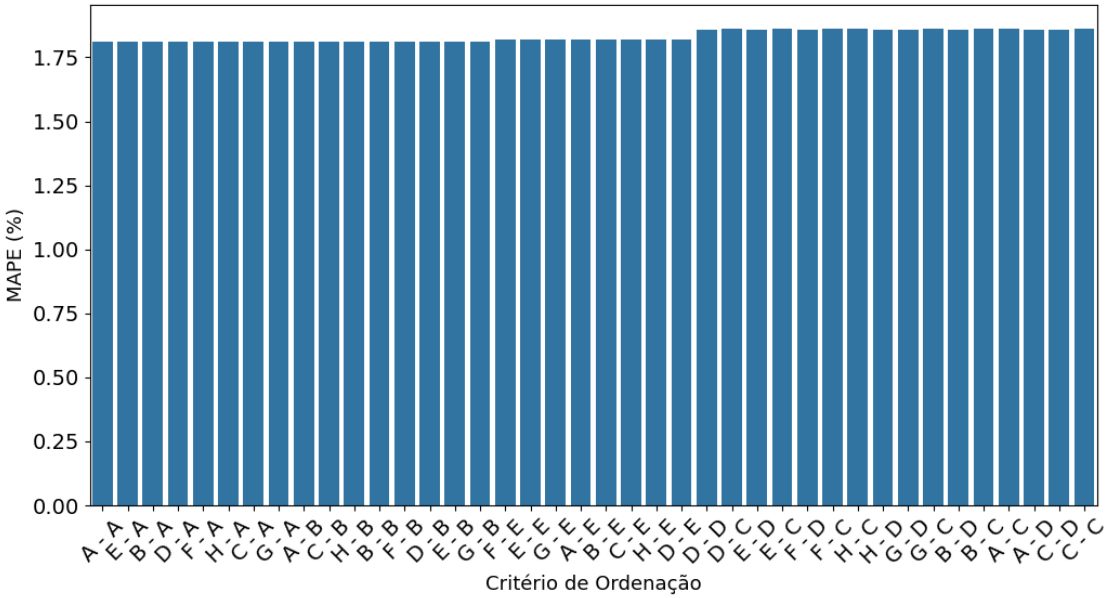


Figura V.66: Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Wine*

Em relação aos dados da base *Wine*, os erros percentuais médios (Figura V.65 entre os critérios de ordenação apresentaram pouca diferença significativa. No entanto, ao analisar os erros mínimos (Figura V.66), foi possível observar diferenças nos critérios de ordenação dos atributos, sendo que *lessCorrelation* (A) e *lessMissing* (B) apresentaram os menores erros mínimos. Em seguida, os critérios *noSort* (E) mostraram um desempenho intermediário, enquanto *moreCorrelation* (C) e *moreMissing* (D) obtiveram os maiores erros mínimos.

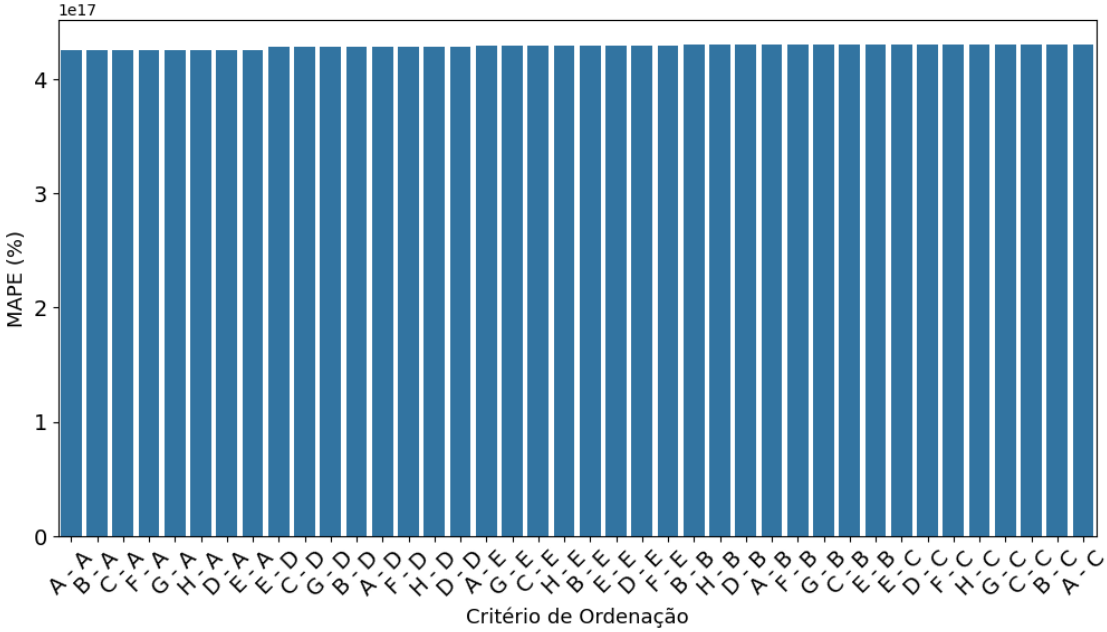


Figura V.67: Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Pima*

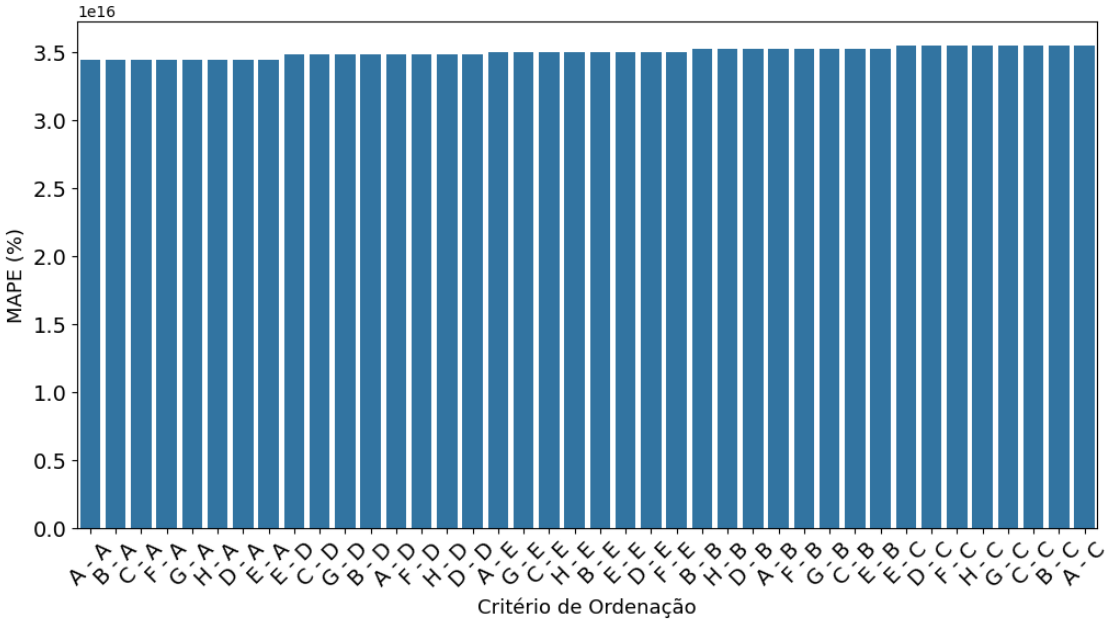


Figura V.68: Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Pima*

Nos dados da base *Pima*, não foi observada uma diferença significativa nos erros médios (Figura V.67). Semelhante ao comportamento observado na base *Wine*, ao analisar os erros mínimos (Figura V.68), verificou-se que a ordem dos atributos teve influência no incremento do erro, indicando que diferentes critérios de ordenação impactam o desempenho de forma notável.

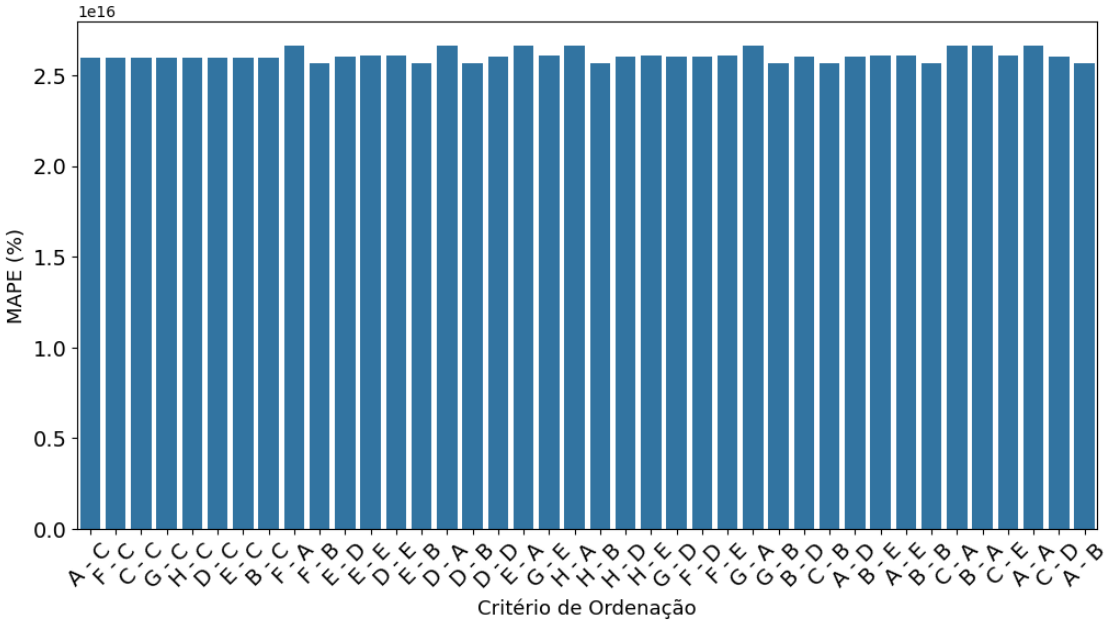


Figura V.69: Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Boston*

Na base *Boston*, observou-se uma pequena diferença nos erros médios (Figura V.69, com os critérios de ordenação dos atributos se classificando de acordo com os seguintes valores de erro médio:

lessMissing (B), *moreCorrelation* (C), *moreMissing* (D), *noSort* (E) e, por fim, *lessCorrelation* (A). Contudo, ao analisar os erros mínimos (Figura V.70), apenas o critério de ordenação *moreCorrelation* (C) apresentou um erro levemente inferior em comparação aos demais critérios.

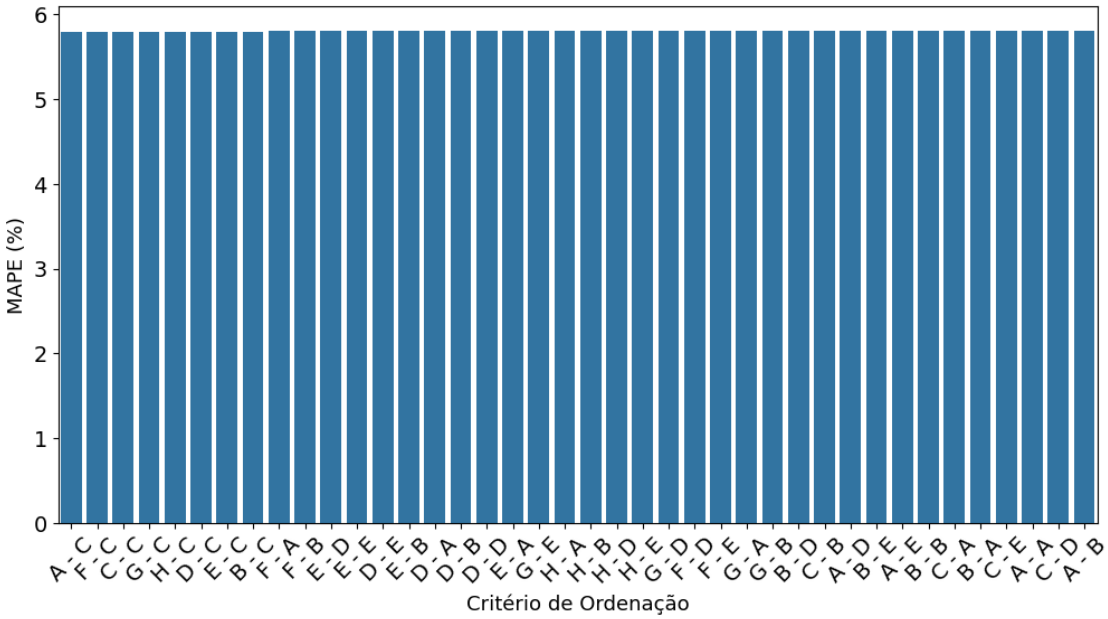


Figura V.70: Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Boston*

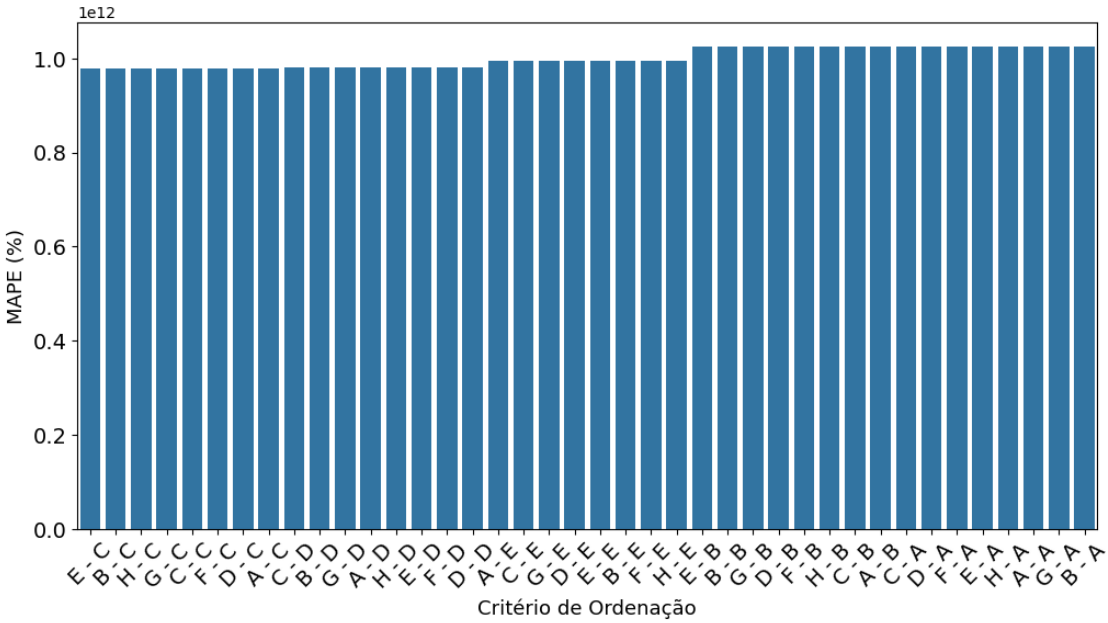


Figura V.71: Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Abalone*

Na base *Abalone*, os erros médios (Figura V.71) e mínimos (Figura V.72) indicaram que a ordem dos atributos impacta no aumento dos erros. A ordem do desempenho dos critérios foi: *moreCorrelation* (C), *moreMissing* (D), *lessMissing* (B), *noSort* (E) e *lessCorrelation* (A).

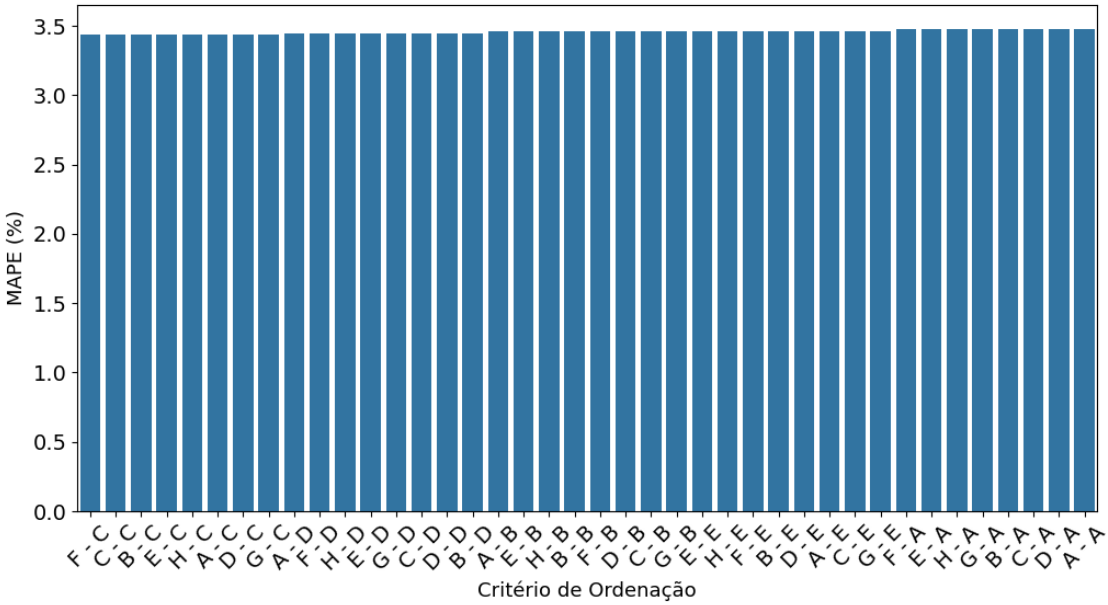


Figura V.72: Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Abalone*

Na base *Gamma*, ao analisar a diferença entre os erros médios (Figura V.73) e os erros mínimos (Figura V.74), observou-se que essa diferença foi insignificante. Contudo, é importante destacar que a amplitude dos erros mínimos foi ligeiramente menor, apresentando uma redução de 0,4 pontos percentuais em relação aos erros médios. Essa pequena diferença sugere que, apesar da ausência de uma variação significativa entre os dois tipos de erro, os erros mínimos apresentaram um desempenho ligeiramente superior quando comparados aos erros médios. Assim, pode-se inferir que a variação nos resultados de imputação não foi substancial entre as duas métricas de erro analisadas.

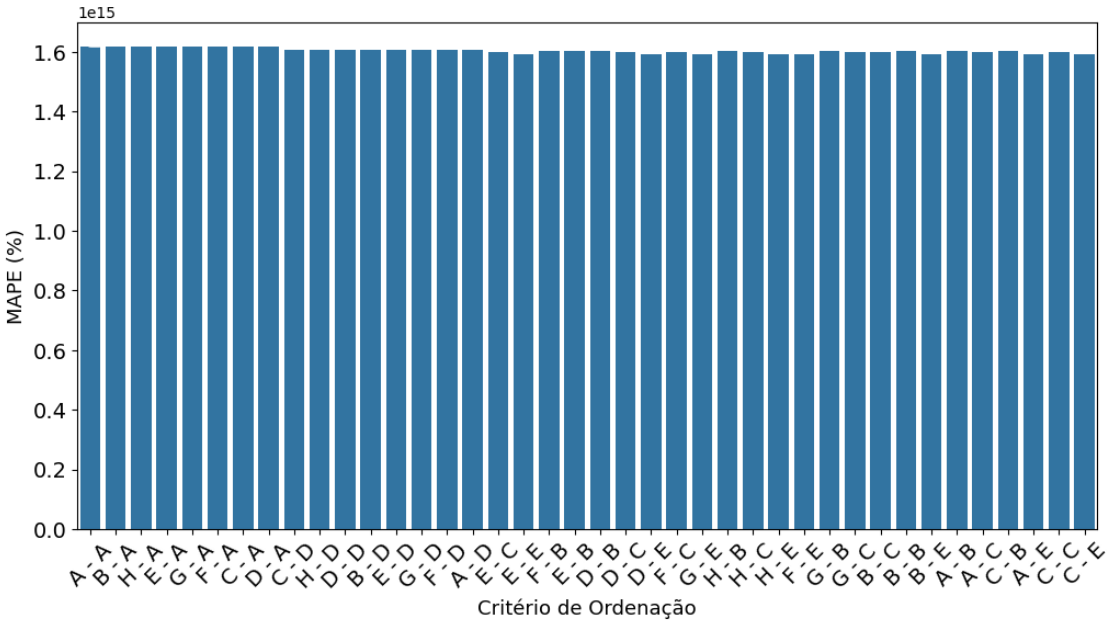


Figura V.73: Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Gamma*

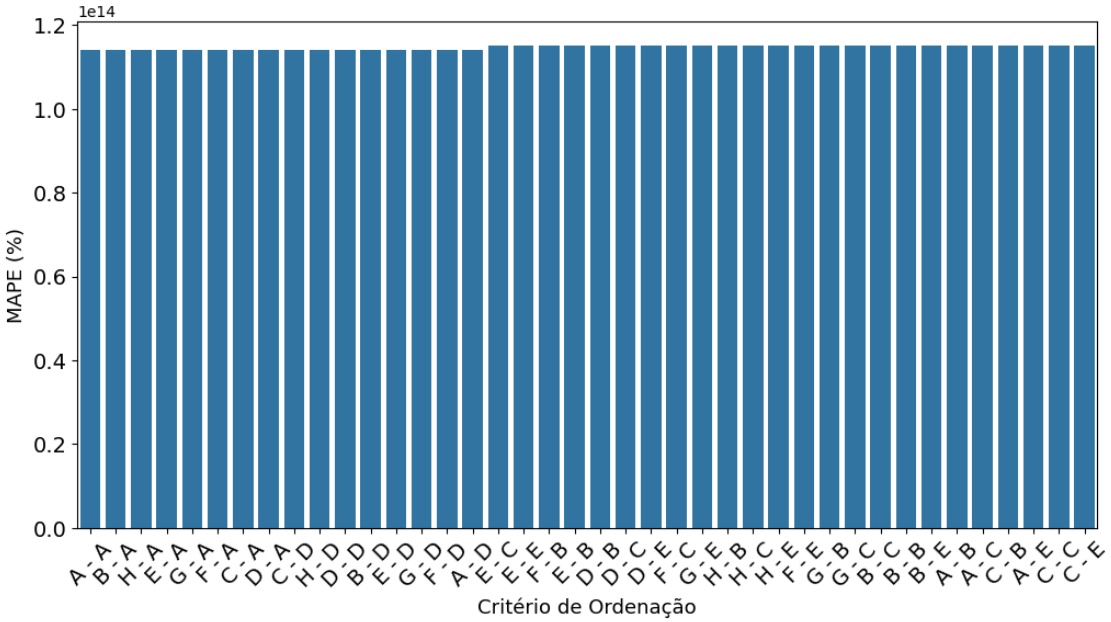


Figura V.74: Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Gamma*

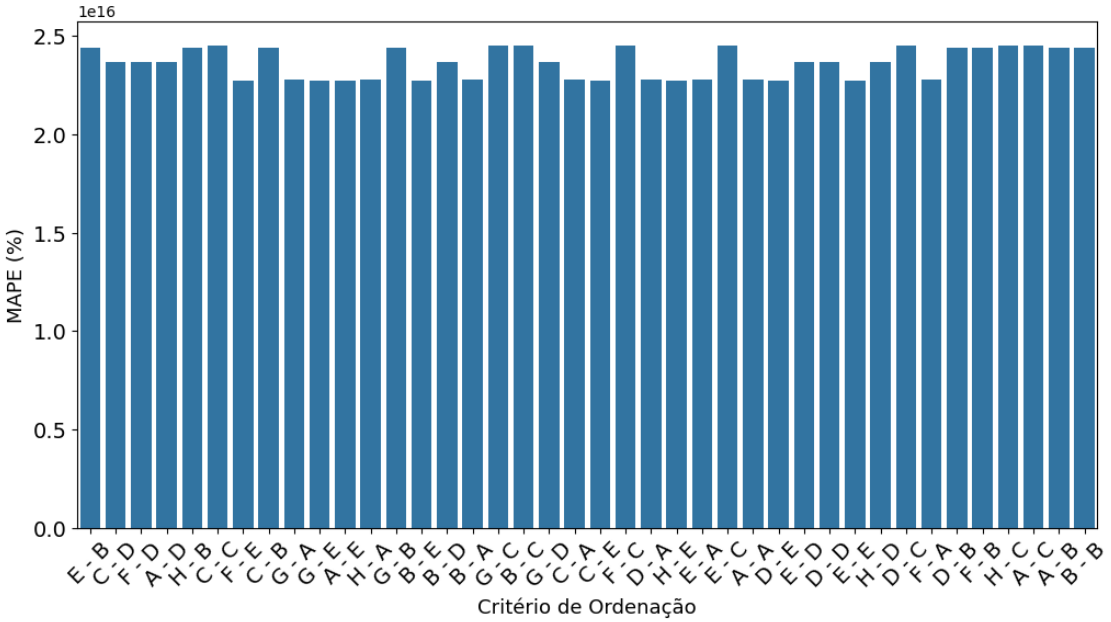


Figura V.75: Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Forest*

Quanto aos dados da base *Forest*, os erros mínimos (Figura V.76) não apresentaram diferença significativa entre os critérios de ordenação, tanto de grupo quanto de atributo. No entanto, ao analisar os erros médios (Figura V.75), observou-se que a ordem dos atributos teve impacto no aumento dos erros. Os critérios de ordenação de atributos que apresentaram o melhor desempenho nesta base foram *lessCorrelation* (A) e o *noSort* (E). Vale destacar que, embora ambos os critérios tenham se destacado, o erro médio foi o mesmo para os dois.

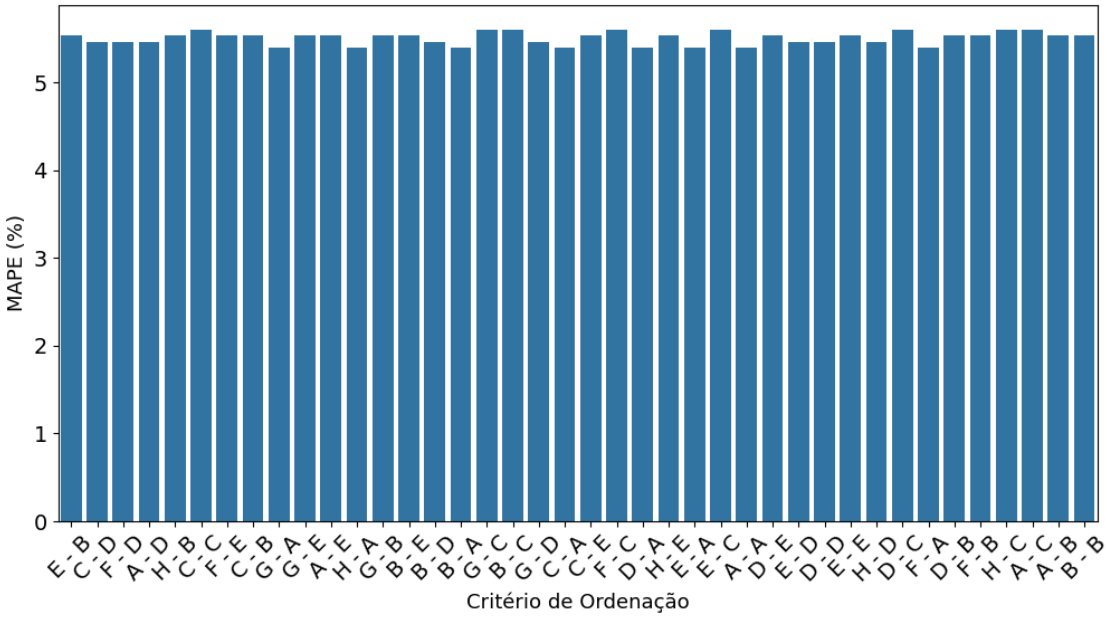


Figura V.76: Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Forest*

Na base *Glass*, os erros mínimos (Figura V.78) não mostraram diferença significativa entre os critérios de ordenação. Contudo, os erros médios (Figura V.77) indicaram que os critérios de ordenação *lessCorrelation* (A), *lessMissing* (B) e *moreMissing* (D) apresentaram um desempenho superior, embora todos tenham exibido resultados idênticos entre si. Em seguida, os critérios *moreCorrelation* (C) e *noSort* (E) também demonstraram resultados semelhantes, embora com um desempenho ligeiramente inferior em comparação aos primeiros.

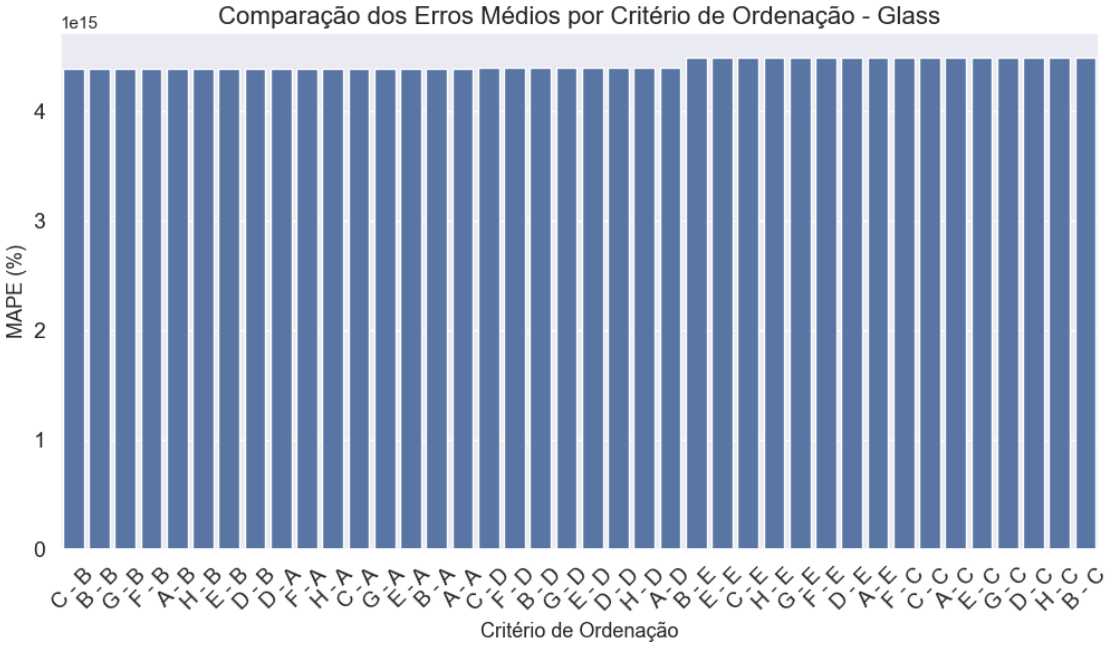


Figura V.77: Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Glass*

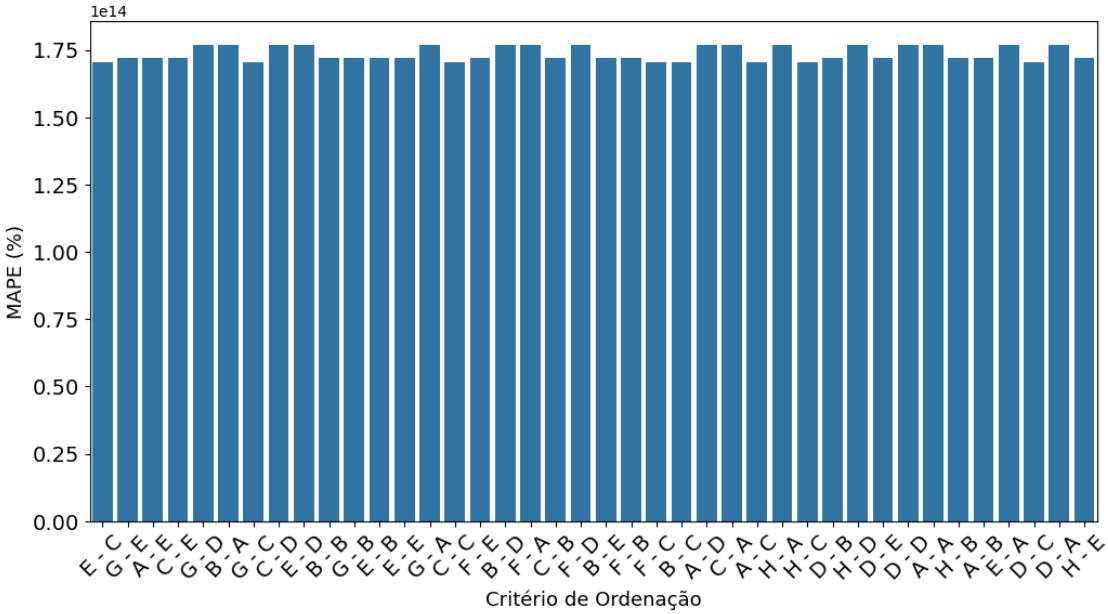


Figura V.78: Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Glass*

Por fim, nos dados da base *Yeast*, os erros mínimos (Figura V.80) não apresentaram diferença significativa entre os critérios de ordenação. Entretanto, ao analisar os erros médios (Figura V.79), foi possível identificar três grupos distintos: o primeiro, com o menor percentual de erro médio, foi composto pelo critério *moreCorrelation* (C); seguido pelo bloco com os critérios *moreMissing* (D), *lessMissing* (B) e *lessCorrelation* (A); e, por fim, o critério *noSort* (E) apresentou os maiores erros médios.

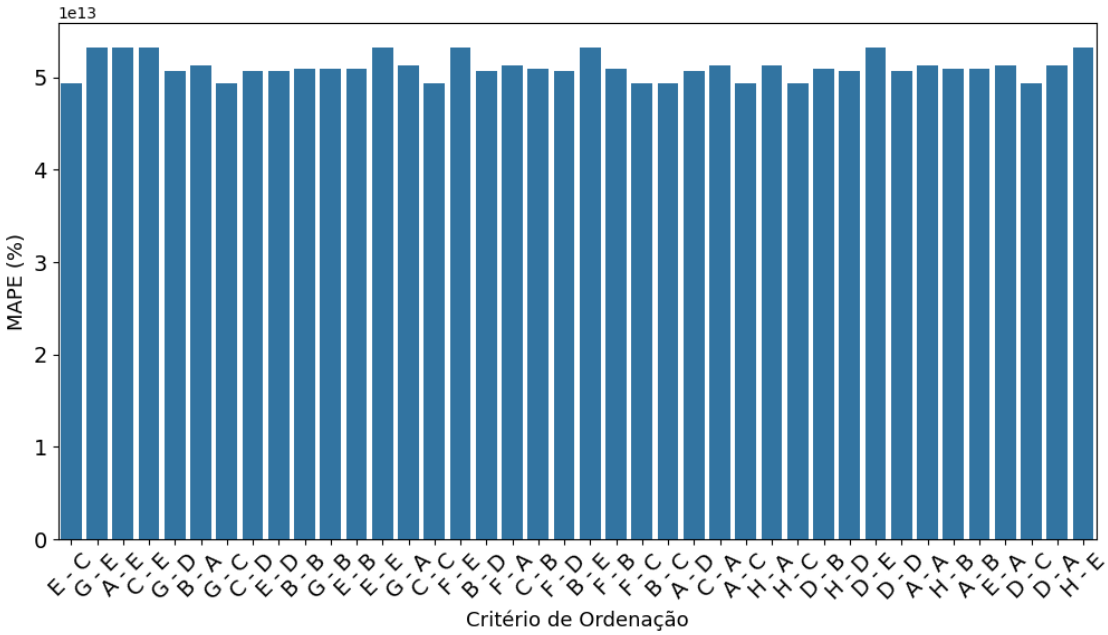


Figura V.79: Comparação dos erros médios de imputação (%) por critério de ordenação nos dados do *Yeast*

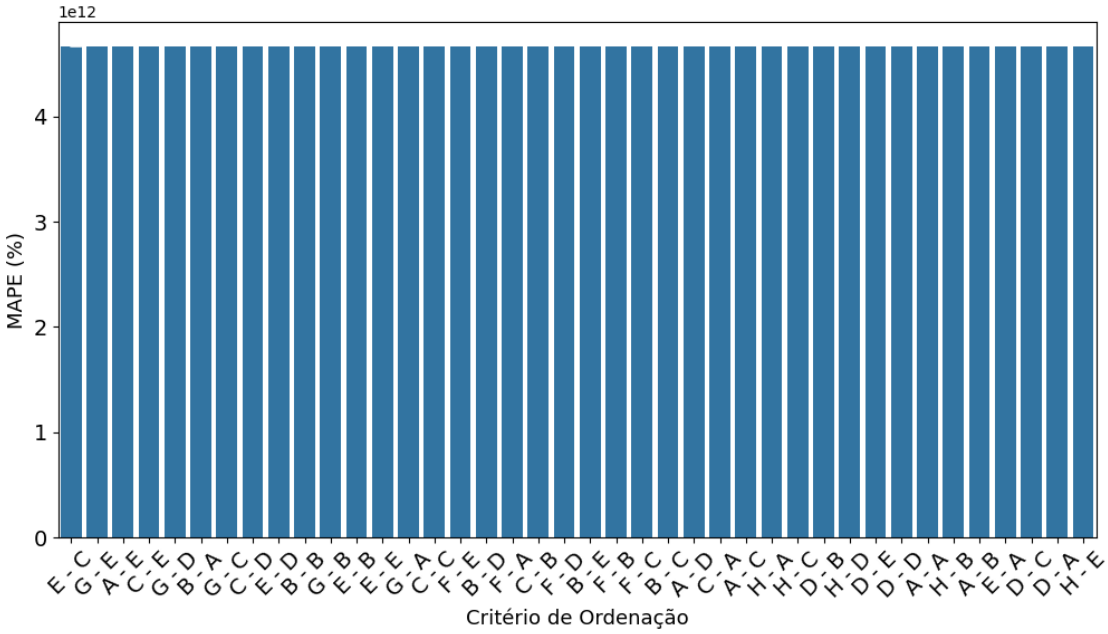


Figura V.80: Comparação dos erros mínimos de imputação (%) por critério de ordenação nos dados do *Yeast*

Em conclusão, a análise comparativa entre os erros médios e mínimos em relação aos critérios de ordenação revelou que não existe uma diferença significativa entre as ordens dos grupos. Entretanto, em muitas bases de dados, a ordem dos atributos teve um impacto mais significativo nos erros mínimos do que nos erros médios. Embora as diferenças nos erros médios fossem geralmente pequenas, especialmente nas bases *Wine*, *Pima* e *Gamma*, os erros mínimos mostraram variações mais pronunciadas, com alguns critérios de ordenação, como *lessCorrelation* e *lessMissing*, apresentando desempenho superior. Nas bases *Boston* e *Abalone*, por exemplo, a ordem dos atributos teve uma influência mais notável, com critérios como *moreCorrelation* e *moreMissing* resultando em menores erros mínimos em determinados contextos.

Além disso, em algumas bases, como *Glass* e *Yeast*, as diferenças entre os erros mínimos foram mais evidentes, sugerindo que, nesses casos, a escolha do critério de ordenação dos atributos pode ter um impacto mais direto no desempenho de imputação, particularmente quando o erro mínimo é o principal critério de avaliação. Por outro lado, os erros médios apresentaram variações mais suaves, evidenciando que a escolha da ordem dos atributos, embora importante, tem um efeito menos pronunciado em termos de média global.

V.4.3 Comparação do desempenho da Imputação em Cascata de acordo com a Classe de Correlação e Outliers das Bases

Um dos diferenciais deste trabalho em relação ao de Ferlin [2008] é que a seleção das bases de dados levou em consideração diferentes níveis de correlação e presença de outliers. Portanto, nesta

seção, avaliaremos os resultados com base nas classes de correlação e outliers, permitindo uma análise mais aprofundada das implicações desses fatores no desempenho dos métodos de imputação.

Ao comparar os erros médios de imputação (Figura V.81), segmentados pela Classe Correlação-Outlier, verificamos que a classe "Médio C – Baixo O" apresentou o menor erro em comparação com as demais classes, com destaque para o método de Imputação em Cascata Simplificada, que obteve o melhor desempenho. Além dessa classe, as classes "Alto C – Baixo O" e "Médio C – Alto O" também se destacaram por apresentarem bons resultados com a Imputação em Cascata Simplificada. Outro ponto importante é que a imputação por mediana obteve o menor erro médio nas classes "Alto C – Médio O", "Baixo C – Alto O", "Baixo C – Médio O" e "Baixo C – Baixo O".

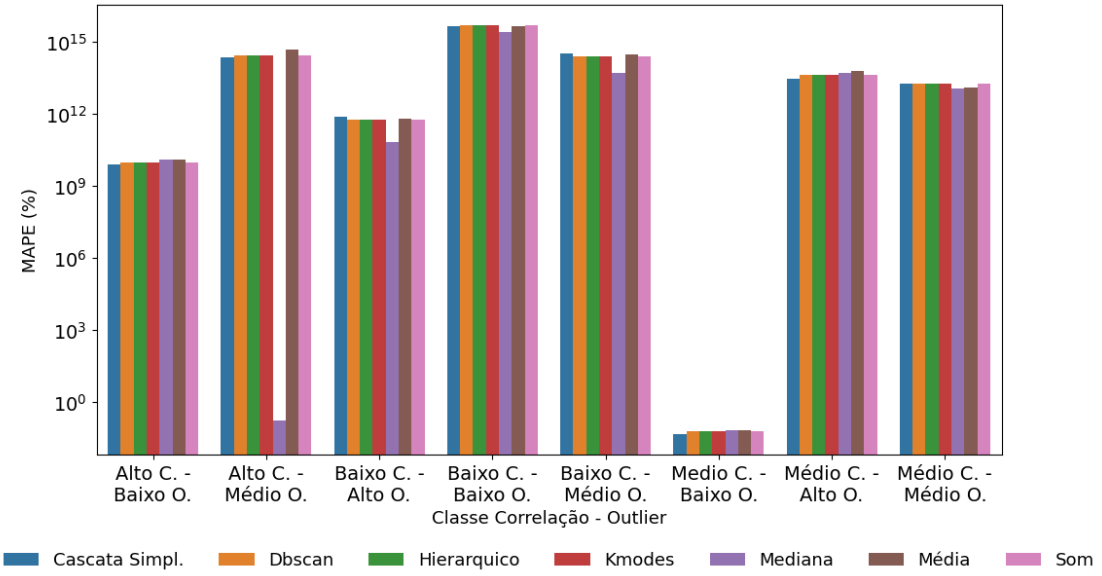


Figura V.81: Comparação dos erros médios de imputação (%) por classe correlação-outliers

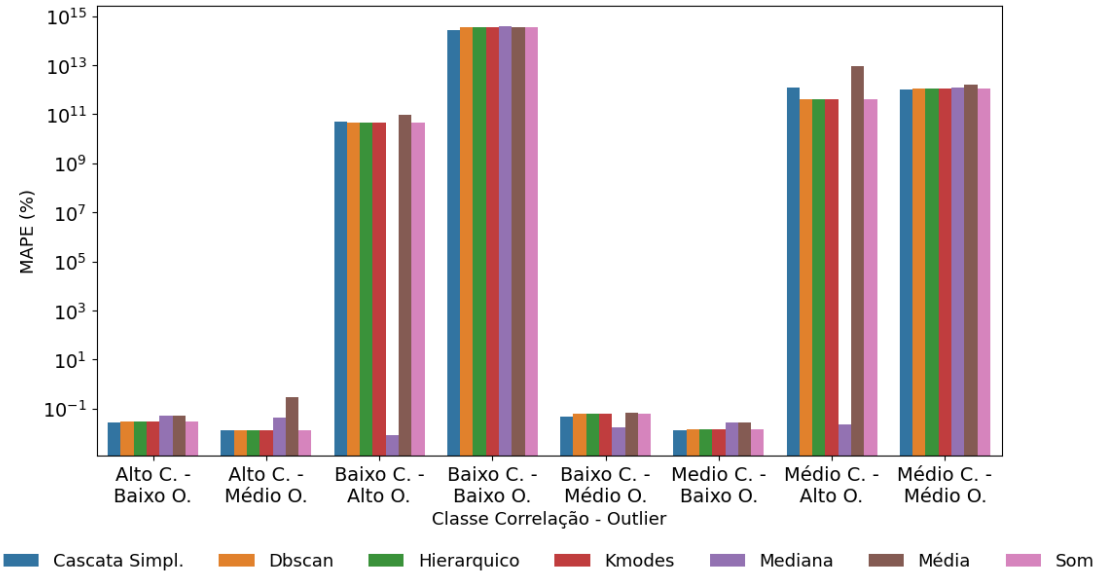


Figura V.82: Comparação dos erros mínimos de imputação (%) por classe correlação-outliers

Já ao analisar os erros mínimos (Figura V.82), observamos uma variação significativa entre os erros mínimos e médios, o que gera mudanças consideráveis nos resultados finais ao considerar a métrica de erro mínimo. Nesse contexto, verificou-se que a classe "Alto C – Médio O" passou a apresentar os piores resultados com os métodos de imputação univariada, enquanto a classe "Baixo C – Baixo O" obteve o melhor desempenho com a Imputação em Cascata Simplificada. Por fim, na classe "Médio C – Alto O", houve uma deterioração no desempenho da Imputação em Cascata Simplificada, sendo superada pela imputação por mediana, que se tornou o método de melhor desempenho nessa classe.



Figura V.83: Comparação dos bias de correlação médio por classe correlação-outliers

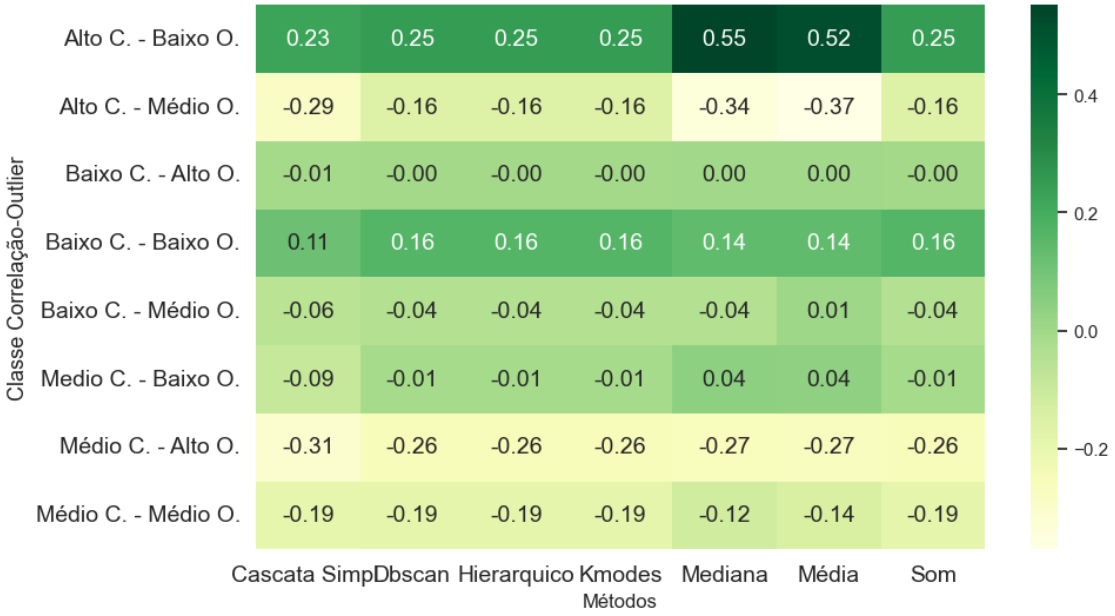


Figura V.84: Comparação dos bias de correlação mínimo por classe correlação-outliers

Comparando o bias de correlação médio (Figura V.83) e o bias de correlação mínimo (Figura V.84), observamos uma variação significativa entre as métricas. No entanto, os métodos que apresentaram a menor alteração em relação à base original mantiveram-se consistentes, com exceção da classe "Alto C – Médio O". No caso do bias de correlação mínimo, destacaram-se as classes "Baixo C – Alto O", "Baixo C – Baixo O", "Baixo C – Médio O" e "Médio C – Baixo O", que apresentaram os melhores resultados.

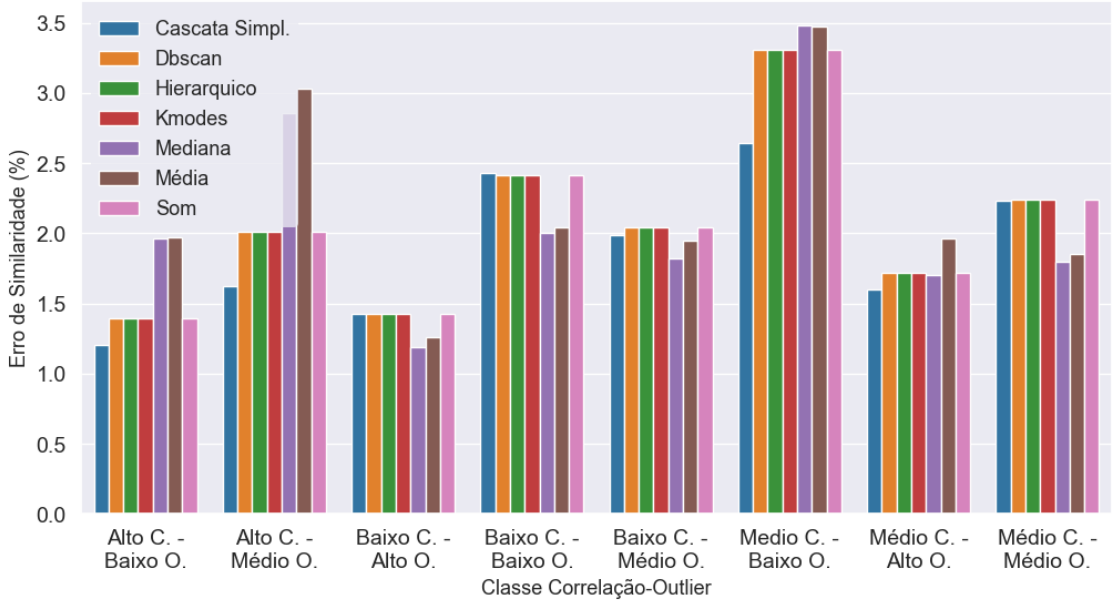


Figura V.85: Comparação dos erros de similaridade médios por classe correlação-outliers

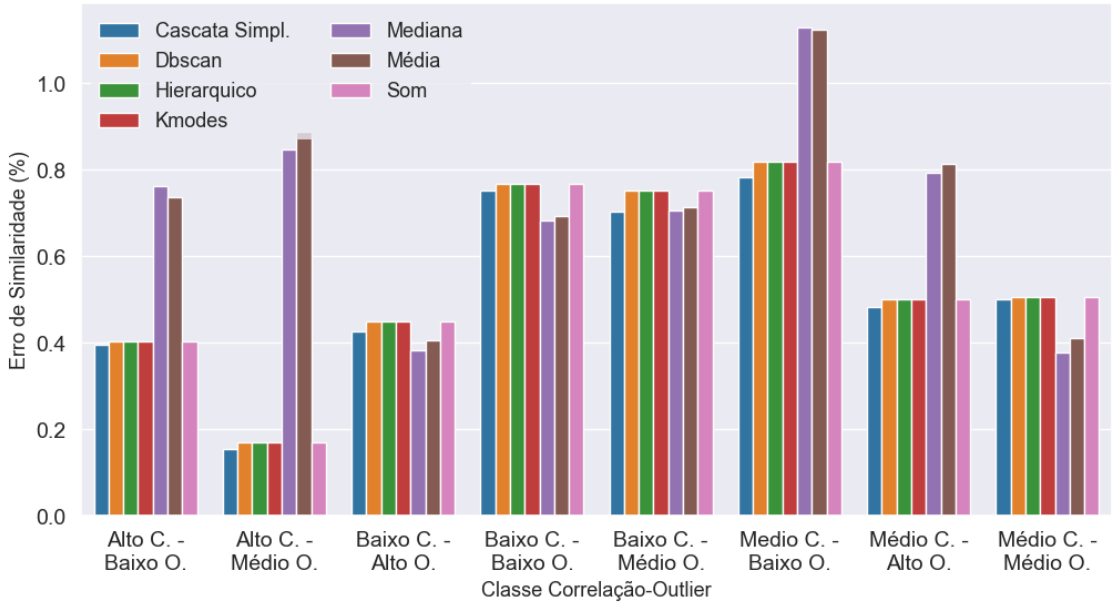


Figura V.86: Comparação dos erros de similaridade mínimos por classe correlação-outliers

Ao analisar os erros de similaridade entre as classes correlação-outlier, verificamos que, apesar da grande variação observada entre o erro médio (Figura V.85) e o erro mínimo (Figura V.86),

alguns padrões se mantiveram. Em particular, as imputações univariadas apresentaram os melhores resultados nas classes "Baixo C – Alto O", "Baixo C – Médio O", "Baixo C – Baixo O" e "Médio C – Alto O". Por outro lado, as classes "Alto C – Baixo O", "Alto C – Médio O", "Médio C – Baixo O" e "Médio C – Alto O" apresentaram a Imputação em Cascata Simplificada como o método de melhor desempenho.

Na Tabela B.1 (Apêndice B) comparamos as configurações dos critérios de ordenações e as classes correlação-outlier, onde foi possível confirmar que a ordem dos grupos não influencia os erros médios de imputação, corroborando a conclusão apresentada na Seção V.4.2, em que se destacou que apenas a ordem dos atributos tem impacto nos resultados. Ao observar as classes de correlação e *outliers*, verificamos que: (i) a classe "Baixo C – Baixo O" obteve o menor erro utilizando a ordem de atributos *lessCorrelation* (A); (ii) as classes "Baixo C – Médio O" e "Médio C – Médio O" apresentaram melhor desempenho com o critério de ordem *noSort* (E); (iii) as classes "Baixo C – Alto O" e "Alto C – Baixo O" se destacaram com a ordem *moreCorrelation*; (iv) e, por fim, as classes "Médio C – Alto O" e "Alto C – Médio O" obtiveram melhores resultados utilizando o critério *lessMissing*. A classe "Médio C – Baixo O" apresentou bons resultados em todos os critérios de ordem dos atributos. Um ponto adicional é que o critério de ordem *moreMissing* não obteve um bom desempenho em nenhuma das classes analisadas.

Em conclusão, a análise comparativa feita nessa Seção revelou os seguintes padrões de desempenho:

- **Classe Baixo C – Baixo O:** Obteve os melhores resultados utilizando a imputação por mediana e a ordem de atributos *lessCorrelation*;
- **Classe Baixo C – Médio O:** Obteve melhor desempenho com o critério de ordem *noSort*;
- **Classe Baixo C – Alto O:** Destacou-se com a ordem de atributos *moreCorrelation*;
- **Classe Médio C – Baixo O:** Obteve bons resultados em todos os critérios de ordem dos atributos, mostrando um desempenho mais equilibrado entre os métodos e as ordens;
- **Classe Médio C – Médio O:** Apresentou os melhores resultados com a ordem de atributos *noSort*;
- **Classe Médio C – Alto O:** Obteve melhores resultados com a Imputação em Cascata Simplificada e a ordem de atributos *lessMissing*;
- **Classe Alto C – Baixo O:** Teve o melhor desempenho com a Imputação em Cascata Simplificada e a ordem *moreCorrelation*;

- **Classe Médio C – Alto O:** Apresentou melhores resultados com a imputação por mediana e a ordem de atributos *lessMissing*.

V.4.4 Comparação do desempenho do valor de k na Imputação em Cascata

Como mencionado anteriormente (Seção IV.4), a imputação em cascata utiliza o algoritmo k-vizinhos mais próximos (k-NN) para identificar os k vizinhos mais próximos de um dado ausente e, a partir disso, preenche o valor ausente com a média dos valores desses vizinhos. Nesta seção, avaliaremos se existe um valor de k que apresenta melhor desempenho, considerando diferentes mecanismos de ausência, classes de correlação-outlier e critérios de ordenação.

Ao comparar o desempenho das métricas em relação aos mecanismos de ausência e ao tamanho de k (Tabela V.10), observou-se que, ao avaliar os erros de imputação, o valor de $k = 1$ apresentou o melhor desempenho na maioria dos mecanismos de ausência, com exceção do MAR 30% e MNAR 10%, onde $k = 5$ teve melhor desempenho, e MCAR 20%, onde $k = 3$ foi mais eficaz. Por outro lado, o desempenho do valor de k em relação ao bias de correlação seguiu uma tendência oposta: todos os mecanismos de ausência apresentaram um bias menor com $k = 10$, exceto nos casos de MCAR 20% e MCAR 30%. Quanto ao erro de similaridade, os melhores resultados foram distribuídos entre os diferentes valores de k: MCAR 10% obteve o melhor desempenho com $k = 1$; $k = 3$ gerou os melhores resultados para MAR 10%, MAR 30% e MCAR 10%; $k = 10$ foi mais eficaz para MCAR 20% e MNAR 10%, enquanto o restante teve melhor desempenho com $k = 5$.

Tabela V.10: Comparativo dos erros de imputação, bias de correlação e erro de similaridade por mecanismos de ausência

Métrica	k	MAR 10%	MAR 20%	MAR 30%	MCAR 10%	MCAR 20%	MCAR 30%	MNAR 10%	MNAR 20%	MNAR 30%
MAPE (e^{13})	1	5,56	74,52	11,59	44,14	87,58	110,11	40,44	65	119,18
	3	5,59	81,61	11,77	44,3	78,79	110,7	40,13	73,68	133,02
	5	5,88	81,78	11,53	49,31	82,01	114,88	39,73	77,13	130,68
	10	6,06	81,23	11,58	46,92	88,22	116,78	43,42	79,36	136,66
Bias de Correl.	1	0,067	0,112	0,128	0,053	0,167	0,263	0,078	0,129	0,311
	3	0,053	0,091	0,128	0,046	0,117	0,26	0,055	0,118	0,278
	5	0,053	0,083	0,119	0,046	0,123	0,261	0,043	0,105	0,276
	10	0,052	0,068	0,115	0,046	0,117	0,258	0,041	0,117	0,282
SIM Erro	1	0,0059	0,0137	0,0193	0,0105	0,0239	0,0401	0,0113	0,0235	0,0399
	3	0,0058	0,0135	0,0188	0,0105	0,0225	0,0383	0,0111	0,0228	0,0379
	5	0,0059	0,0134	0,019	0,0106	0,0223	0,0382	0,0109	0,0226	0,0377
	10	0,006	0,0135	0,0196	0,0107	0,0222	0,0383	0,0109	0,0229	0,0381

Para tornar a apresentação mais compacta, as configurações das classes de correlação-outlier foram codificadas e apresentadas na Tabela V.11. Esta tabela servirá como base para a análise das métricas por classes de correlação-outlier a seguir.

Tabela V.11: Descritivo das configurações das classes correlação-outlier

Classe Correlação	Classe Outlier	Código Classe Correl-Outlier
Baixo	Baixo	0 - 0
	Médio	0 - 1
	Alto	0 - 2
Médio	Baixo	1 - 0
	Médio	1 - 1
	Alto	2 - 2
Alto	Baixo	2 - 0
	Médio	2 - 1

Na análise do impacto de k nas classes de correlação-outlier (Tabela V.12), os melhores resultados de erro de imputação foram observados para $k = 1$ em quase todas as classes, exceto nas classes "Baixo C. – Alto O." (0 - 2) e "Médio C. – Médio O." (1 - 1). As classes "Baixo C. – Baixo O." (0 - 0), "Baixo C. – Alto O." (0 - 2), "Médio C. – Médio O." (1 - 1) e "Alto C. – Baixo O." (2 - 0) apresentaram os melhores resultados tanto no bias de correlação quanto no erro de similaridade com $k = 10$. O mesmo ocorreu na classe "Médio C. – Alto O." (1 - 2), mas nesse caso, com $k = 3$.

Tabela V.12: Comparativo dos erros de imputação, bias de correlação e erro de similaridade por mecanismos de ausência

Métricas	k	0 - 0	0 - 1	0 - 2	1 - 0	1 - 1	1 - 2	2 - 0	2 - 1
MAPE (e^{11})	1	45638,59	1271,8	5,93	0	213,98	339,34	0,09	2139,89
	3	45801	2794,76	4,54	0	175,45	368,71	0,092	2375,61
	5	46087,92	3007,92	5,17	0	165,3	390,41	0,092	3048,11
	10	47133,46	3306,75	7,48	0	155,3	484,94	0,094	3155,29
Bias de Correl.	1	32,85	13,01	4,71	26,08	-10,54	-15,73	71,54	-5,58
	3	26,55	10,73	3,61	21,34	-9,14	-14,41	68,33	-5,26
	5	24,95	11,08	3,26	20,57	-8,7	-14,47	67,26	-5,26
	10	24,55	11,71	3,02	20,61	-8,44	-14,83	66,19	-5,37
SIM Erro	1	2,73	2,13	1,6	3,13	2,4	1,77	1,36	1,6
	3	2,41	2,03	1,41	3,14	2,23	1,63	1,36	1,9
	5	2,3	2,01	1,37	3,17	2,18	1,64	1,36	2,03
	10	2,21	1,96	1,34	3,25	2,14	1,74	1,35	2,2

Embora não tenha sido apresentada uma tabela devido ao seu tamanho (41 linhas × 13 colunas), avaliamos o valor de k nos critérios de ordenação. Em todas as configurações de ordem de grupo e de atributo testadas, sem exceção, os melhores resultados no erro de imputação foram obtidos com $k = 1$, enquanto no bias de correlação o melhor desempenho foi registrado com $k = 10$. No caso do erro de similaridade, os melhores resultados foram alcançados com $k = 5$.

V.4.5 Comparação do desempenho da Imputação em Cascata usando dados normalizados

Neste estudo, optou-se por não normalizar os dados, a fim de avaliar a eficiência das técnicas na presença de diferentes graus de outliers, divergindo da abordagem adotada por Ferlin [2008] (Seção III.2.1). No entanto, nesta seção, realizaremos uma comparação simples entre o desempenho da Imputação em Cascata em dados originais e normalizados. Para essa análise, utilizaremos a base *Wine*, que já foi empregada neste trabalho, juntamente com a base *Iris*. A escolha da base *Iris* deve-se ao fato de Ferlin [2008] também ter utilizado esses dados em seu estudo, além de ser uma base pequena e amplamente utilizada no meio acadêmico.

Observamos que na base *Wine Normalizada* há uma variação maior entre o erro médio de imputação (Figura V.87) e o erro mínimo de imputação (Figura V.88), especialmente nos mecanismos de ausência MAR 20% e MNAR 20%. Também é possível perceber que as imputações por média e/ou mediana se destacaram como os melhores métodos nos mecanismos de ausência MAR 10%, MAR 20%, MCAR 10% e MNAR 10%, tanto nos erros médios (Figura V.87) quanto nos erros mínimos (Figura V.88).

Ao comparar os resultados da base *Wine* original e da base *Wine Normalizada*, verificamos que a escala dos erros aumentou consideravelmente; enquanto os gráficos da base original apresentavam erros de até 14%, com os dados normalizados, os erros chegaram a quase 140%. Além disso, observou-se que, nos dados originais da *Wine* (Figura V.25 e V.26), o método de Imputação em Cascata Simplificada obteve o melhor desempenho em todos os mecanismos de ausência, com exceção de MCAR 30% e MNAR 30%.

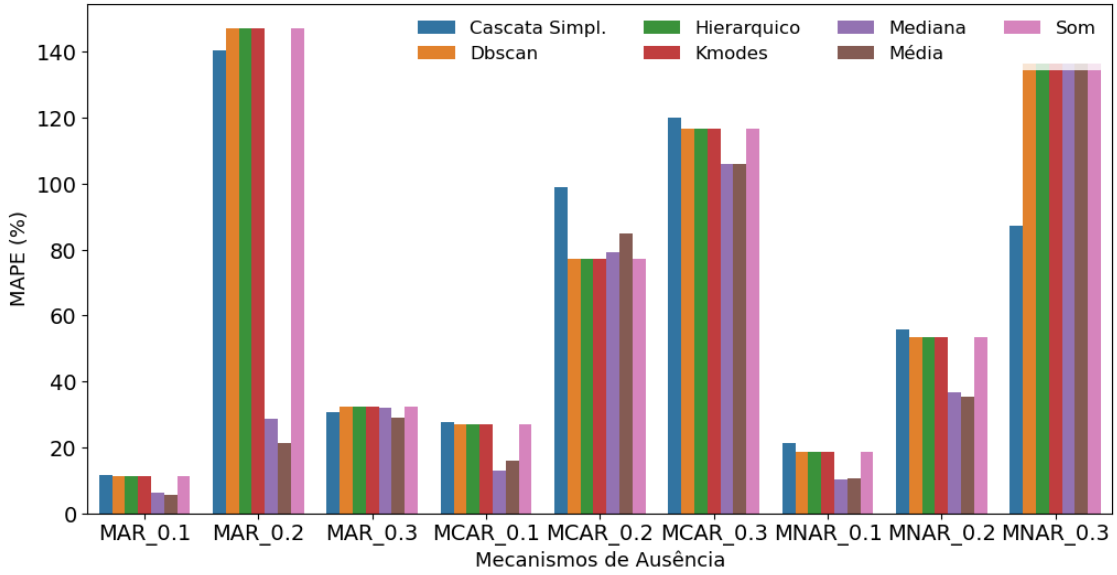


Figura V.87: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Wine Normalizado*

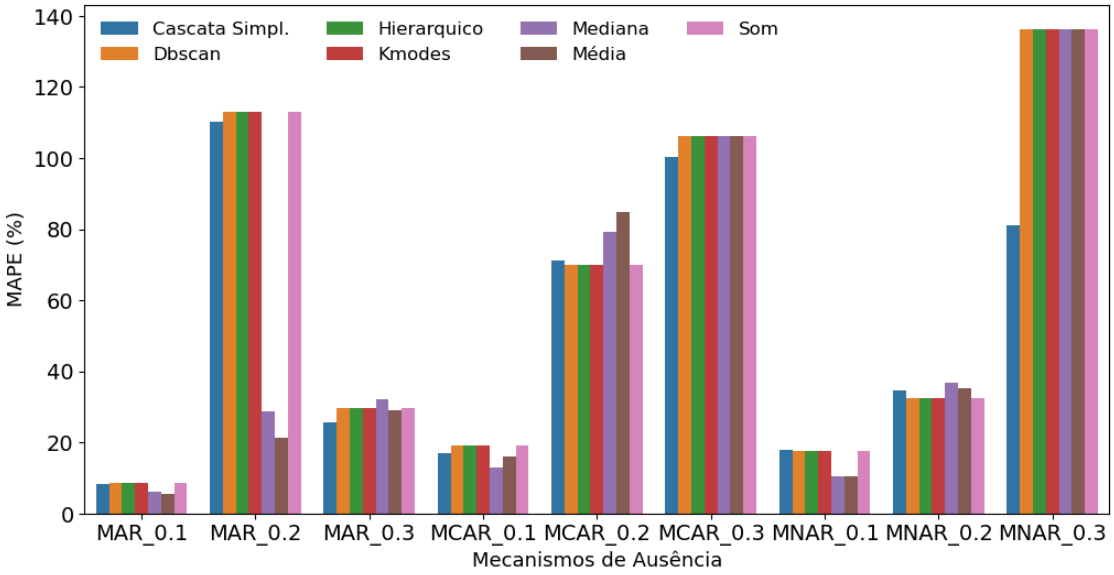


Figura V.88: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Wine Normalizado*

Nos dados da base *Iris*, a diferença entre os erros médios (Figura V.89) e mínimos (Figura V.90) de imputação foi pouco significativa, ou seja, a distribuição dos erros foi praticamente a mesma. Verificou-se também que, em todos os mecanismos de ausência, as imputações univariadas (média e mediana) apresentaram os piores desempenhos.

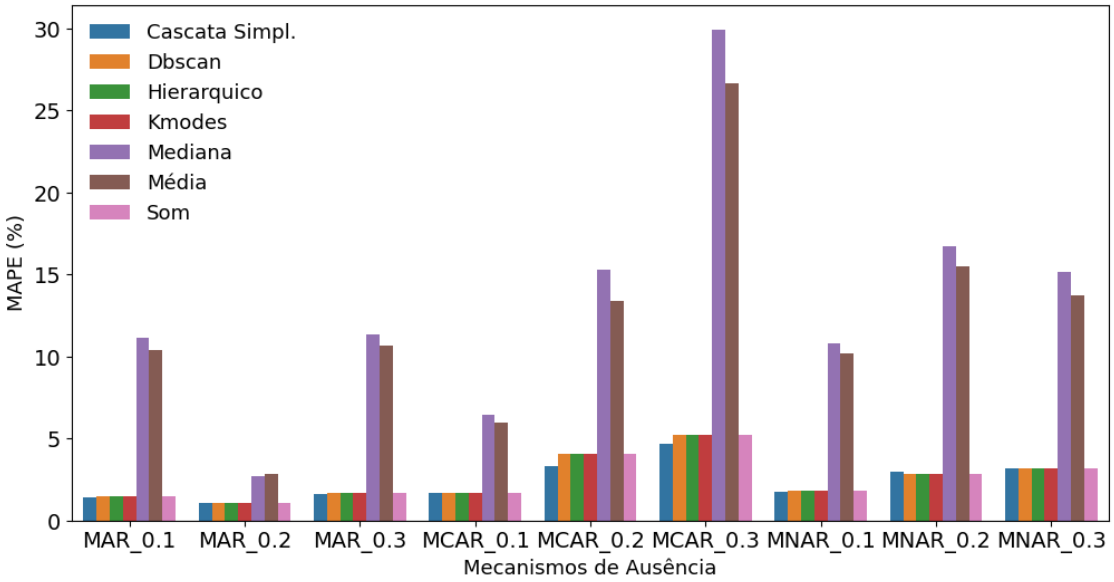


Figura V.89: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Iris*

Entretanto, ao analisar os erros médios (Figura V.91) e mínimos (Figura V.92) de imputação da base *Iris Normalizada*, observou-se uma variação considerável entre os tipos de erro, embora a distribuição entre eles tenha-se mantido. Notou-se que, ao contrário dos dados não normalizados, os métodos de imputação univariada, como média e mediana, tiveram o melhor desempenho em

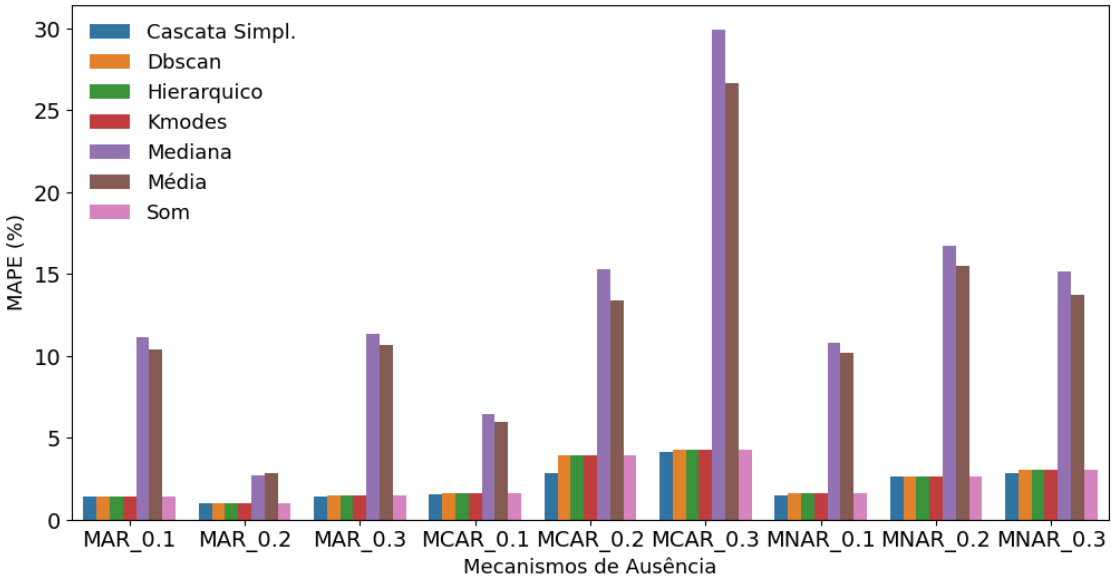


Figura V.90: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Iris*

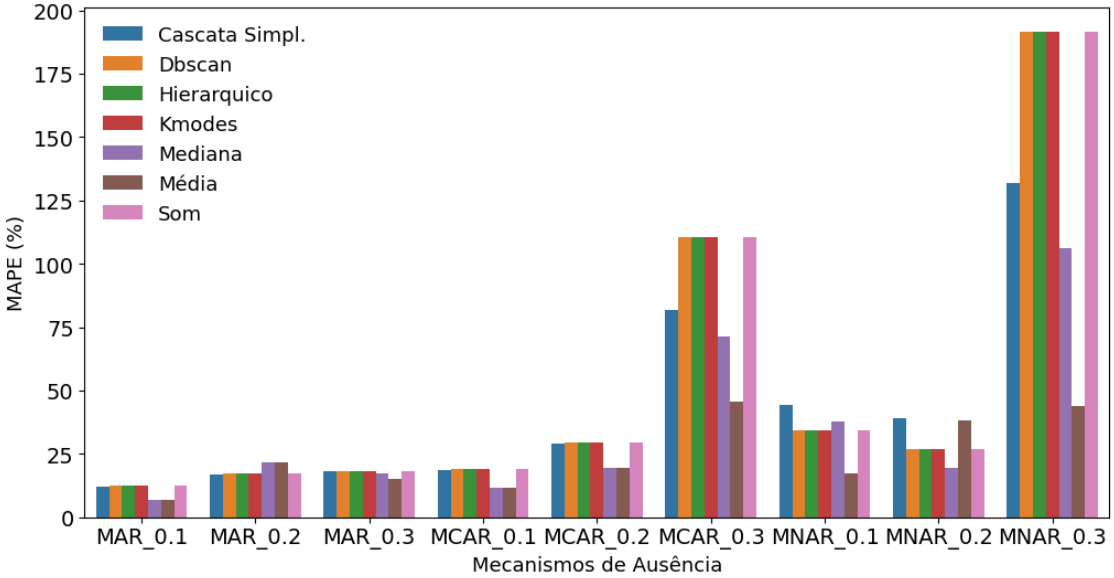


Figura V.91: Comparação dos erros médios de imputação (%) por método de imputação nos dados do *Iris Normalizado*

praticamente todos os mecanismos de ausência, com exceção do MAR 20%. Esse comportamento foi bastante diferente do observado na base *Iris* não normalizada, onde a média e mediana apresentaram os piores resultados em todos os mecanismos de ausência. Também se repetiu o aumento na escala dos erros, que inicialmente era de até 30% na base *Iris* e, após a normalização, chegou a pelo menos 140%.

Ao analisar o bias de correlação nos dados da base *Wine Normalizado*, observamos várias diferenças significativas em relação aos resultados da base *Wine*. Os valores do bias de correlação foram mais elevados na base *Wine Normalizado* (Figura V.93 e V.94) em comparação com os dados

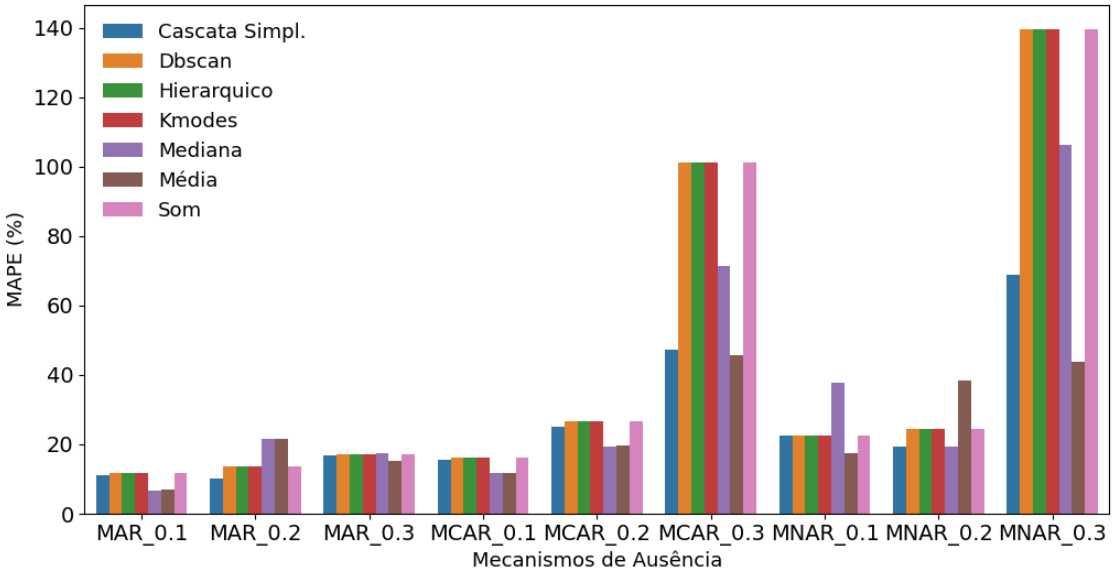


Figura V.92: Comparação dos erros mínimos de imputação (%) por método de imputação nos dados do *Iris Normalizado*

originais da base *Wine* (Figura V.41 e V.42), com exceção do mecanismo de ausência MNAR 30%. Nos dados normalizados da *Wine*, considerando a métrica de bias médio (Figura V.93), obteve-se um bom desempenho nos mecanismos de ausência MAR 10%, MAR 20% e MCAR 10%. Nestes cenários, os métodos que apresentaram os melhores resultados foram DBSCAN, Hierárquico, K-modes e SOM. Nos demais casos, a Imputação em Cascata Simplificada obteve o menor bias de correlação.

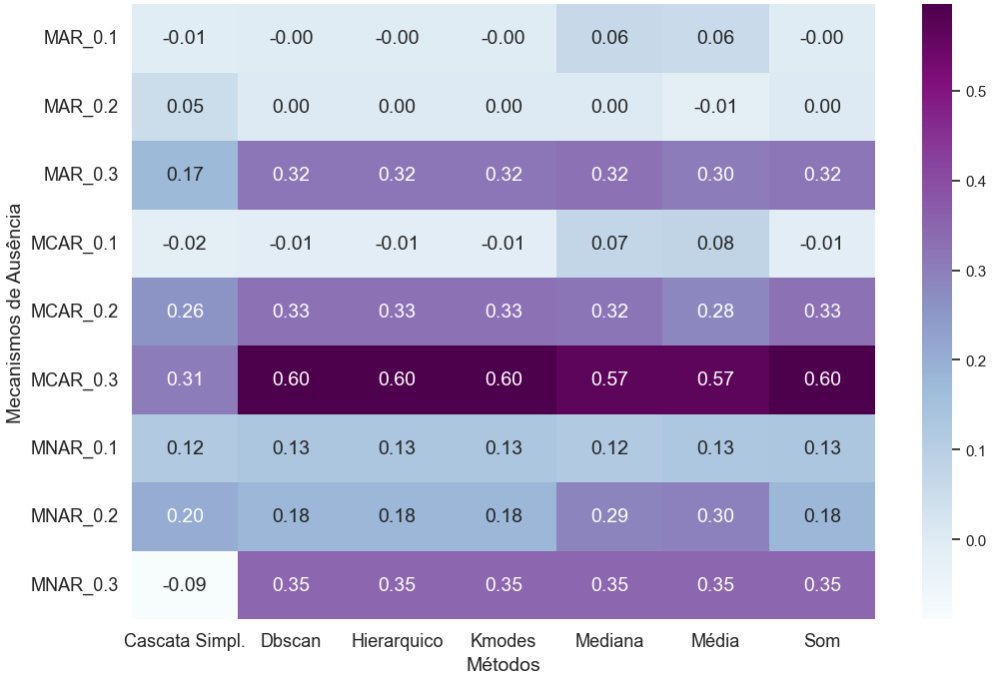


Figura V.93: Comparação do bias de correlação médio entre os dados originais de *Wine Normalizada* e os dados imputados

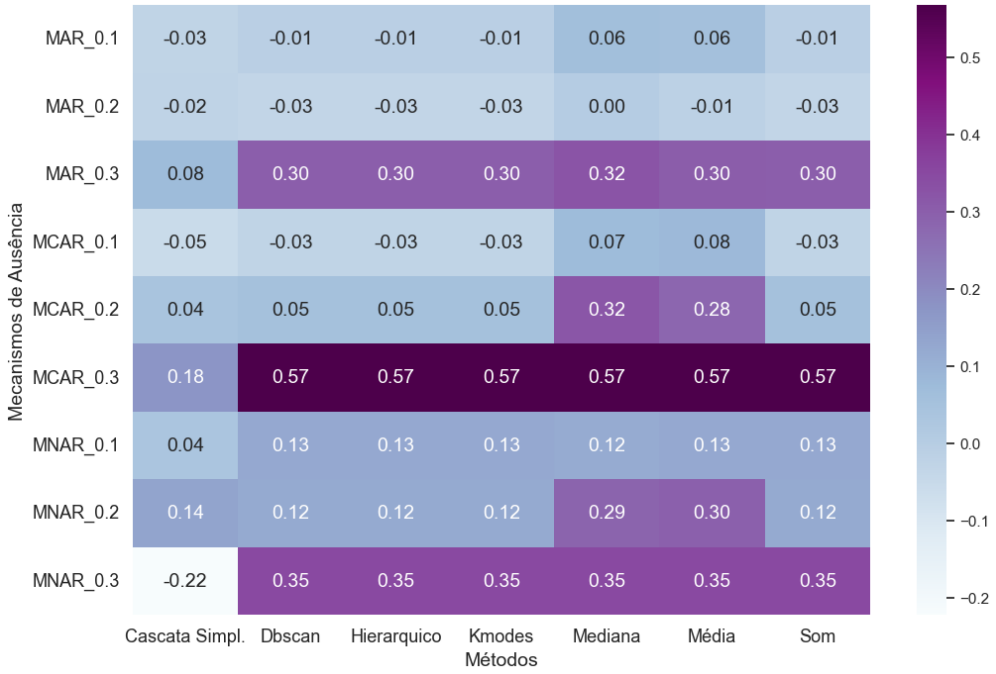


Figura V.94: Comparação do bias de correlação mínimo entre os dados originais de *Wine Normalizada* e os dados imputados

Na base *Iris*, observou-se que a distribuição do bias médio (Figura V.95) e do bias mínimo (Figura V.96) foi consistente. Nestes dados, os valores de bias de correlação foram mais elevados nos mecanismos MCAR 20%, MCAR 30% e MNAR 30%, com a Imputação em Cascata Simplificada se destacando como o melhor método.

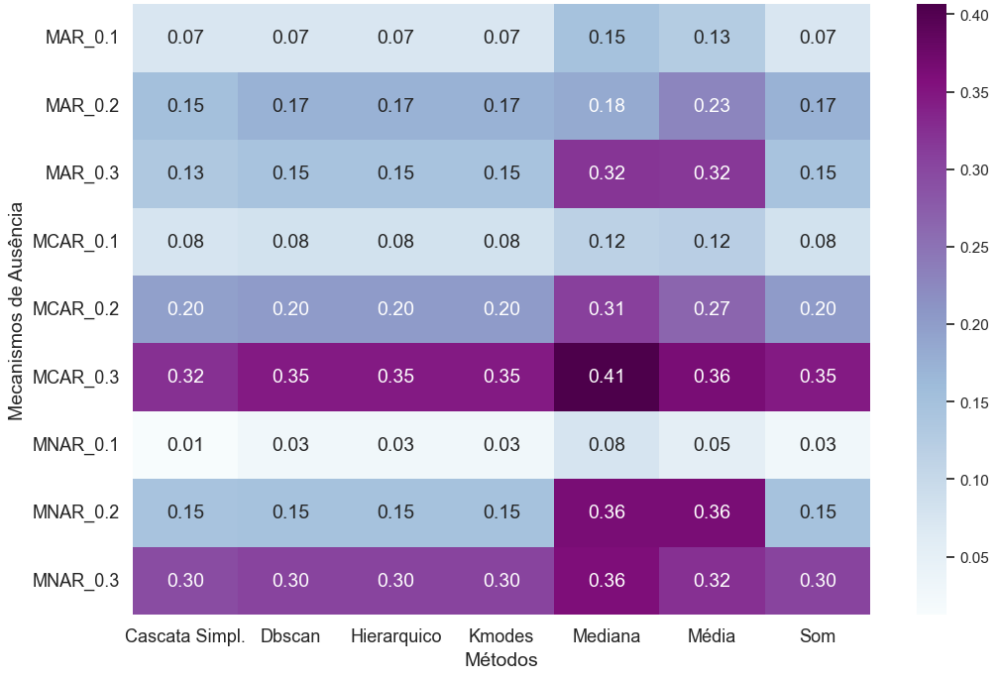


Figura V.95: Comparação do bias de correlação médio entre os dados originais de *Iris* e os dados imputados

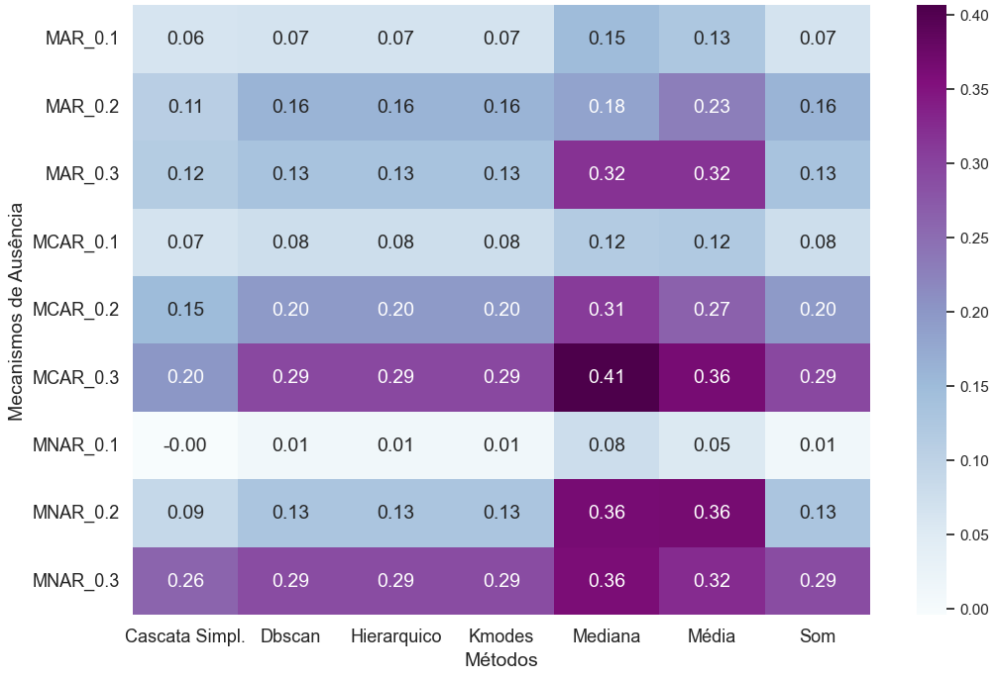


Figura V.96: Comparação do bias de correlação mínimo entre os dados originais de *Iris* e os dados imputados

Na base *Iris Normalizado*, também não foi observada alteração significativa entre o bias médio (Figura V.97) e o bias mínimo (Figura V.98), ou seja, a distribuição dos dados foi mantida. Esse comportamento foi semelhante ao observado na base *Iris*, onde os resultados também indicaram uma estabilidade na distribuição dos dados entre os diferentes tipos de bias.

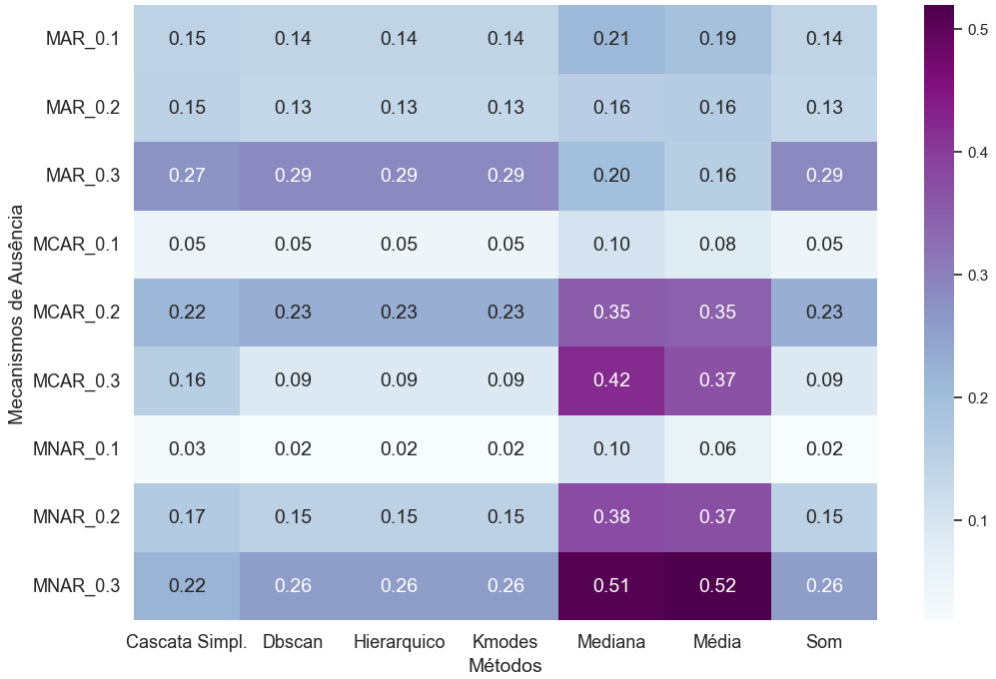


Figura V.97: Comparação do bias de correlação médio entre os dados originais de *Iris Normalizada* e os dados imputados



Figura V.98: Comparação do bias de correlação mínimo entre os dados originais de *Iris Normalizada* e os dados imputados

Quando comparamos os resultados entre as bases *Iris* e *Iris Normalizado*, observamos que, ao normalizar os dados, o bias de correlação nos mecanismos MCAR 30% e MNAR 30% aumentou. No entanto, no caso dos mecanismos MAR 10%, MAR 20% e MAR 30%, houve uma diminuição no bias de correlação (Figura V.95 e Figura V.97). Nos dados da *Iris Normalizado*, os melhores resultados de bias de correlação foram observados nos mecanismos MCAR 10% e MNAR 10% (Figura V.97), onde os métodos de Imputação em Cascata Simplificada, DBSCAN, Hierárquico, K-modes e SOM se destacaram. O mecanismo MCAR 30% apresentou um bom desempenho, com exceção do método de Imputação em Cascata Simplificada.

Não foi observada uma diferença significativa entre os erros de similaridade médio e o erro de similaridade mínimo nos dados *Wine Normalizados*. Devido a isso, só iremos apresentar o erro de similaridade médio (Figura V.99).

Finalmente, ao comparar o erro de similaridade nos dados *Wine Normalizados* (Figura V.99) com os dados originais da *Wine* (Figura V.57), observa-se que os erros dos métodos univariados, como média e mediana, se acentuaram nos mecanismos de ausência que já apresentavam desempenho ruim na base *Wine*. Nos mecanismos MCAR 30% e MNAR 30%, houve uma diminuição no percentual do erro de similaridade para os métodos DBSCAN, Hierárquico, K-modes, mediana, média e SOM, mas, mesmo assim, esses métodos não apresentaram um bom desempenho nesses mecanismos de ausência.

Ao analisar os erros de similaridade na base *Iris* e *Iris Normalizado*, constatou-se que não houve diferença significativa entre os erros médios e mínimos; logo, só serão apresentados aqui os erros de

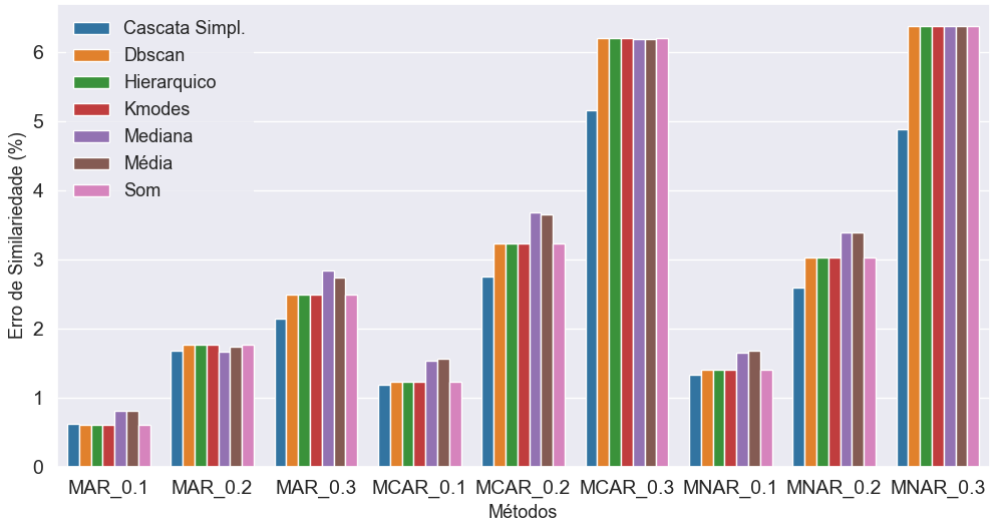


Figura V.99: Comparação da similaridade média entre a base original e a imputada na base *Wine Normalizada*

similaridade médios (Figura V.100 e Figura V.101).

Na base *Iris* observamos que os piores desempenhos foram obtidos pelos métodos de imputação univariada (média e mediana), enquanto a Imputação em Cascata Simplificada obteve bons resultados nos mecanismos de ausência MAR 30%, MCAR 20%, MCAR 30%, MNAR 10% e MNAR 30% (Figura V.100).

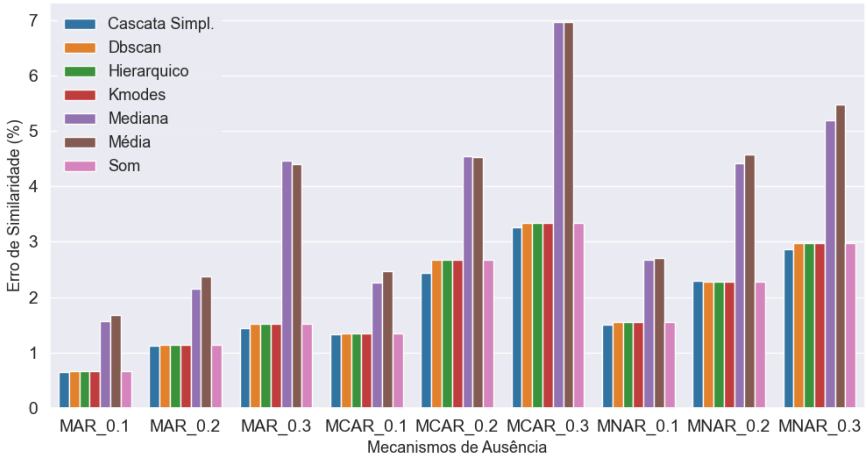


Figura V.100: Comparação da similaridade média entre a base original e a imputada na base *Iris*

Ao comparar os resultados obtidos com os dados *Iris Normalizados* e os dados originais *Iris*, verificamos que a distribuição dos erros de similaridade não apresentou diferenças significativas. Ou seja, os métodos que apresentaram bons resultados na base *Iris* (Figura V.100) também obtiveram bom desempenho na base *Iris Normalizada* (Figura V.101). A principal diferença entre as duas bases foi observada em alguns mecanismos de ausência, como MAR 30%, MCAR 30% e MNAR 30%, onde o percentual de erro diminuiu nos métodos média e mediana, enquanto aumentou nos demais métodos, especialmente no MCAR 30% e MNAR 30%.

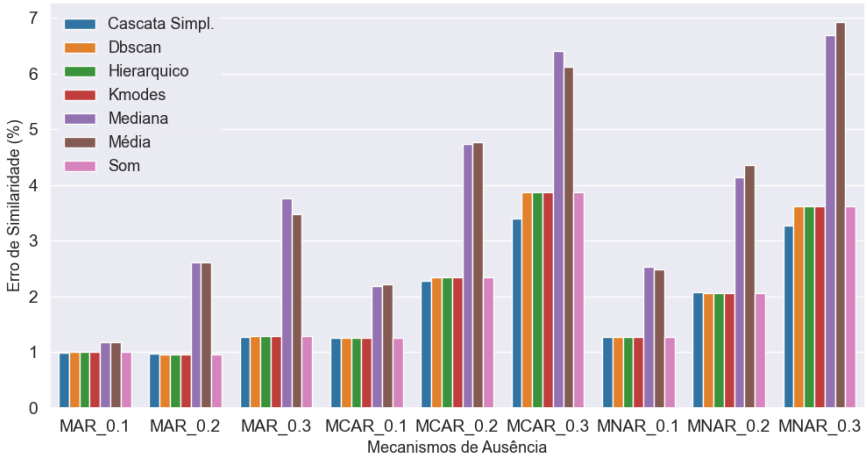


Figura V.101: Comparação da similaridade média entre a base original e a imputada na base *Iris Normalizada*

Em conclusão, a comparação entre as bases *Wine*, *Wine Normalizado*, *Iris* e *Iris Normalizado* revelou importantes diferenças no desempenho dos métodos de imputação. Para a base *Wine*, a normalização dos dados resultou em um aumento significativo nos erros de imputação, com destaque para os métodos univariados, como média e mediana, que apresentaram desempenho inferior em comparação com a Imputação em Cascata Simplificada, que foi mais eficaz nos dados originais. Já para a base *Iris*, a normalização não trouxe grandes mudanças nos erros de similaridade, mas permitiu uma diminuição no bias de correlação em alguns mecanismos de ausência, como MCAR 30% e MNAR 30%. Contudo, a normalização não foi benéfica para todos os cenários, como nos casos de MAR 10%, MAR 30% e MCAR 20%, onde o bias de correlação aumentou. Em termos de erros de similaridade, tanto na base *Wine* quanto na base *Iris*, os métodos de imputação univariada tiveram um desempenho mais equilibrado quando os dados foram normalizados, com a Imputação em Cascata Simplificada se destacando nos cenários com maior presença de correlação e menor presença de outliers. Esses resultados destacam a importância de avaliar as características dos dados ao escolher o método de imputação e a necessidade de considerar os efeitos da normalização na performance dos métodos em diferentes cenários.

V.4.6 Análise da qualidade da imputação

Ao longo deste capítulo, foram avaliadas diversas configurações e o desempenho de cada uma delas. Nesta seção, analisaremos as configurações que apresentaram os menores erros de imputação, comparando seus resultados com os dados originais. As configurações que se destacaram em termos de desempenho, por base e pelas métricas avaliadas, foram:

- *Abalone*
  - MAPE: mecanismos de ausência = MAR 20%; k do kNN = 1; ordem das colunas =

moreCorrelation; método = Cascata Simplificada.

- Bias de Correlação: mecanismos de ausência = MAR 10%; k do kNN = 10; ordem das colunas = *noSort*; método = Cascata Simplificada.
- Erro de Similaridade: mecanismos de ausência = MAR 10%; k do kNN = 1; ordem das colunas = *noSort*; método = Cascata Simplificada

- *Boston*
  - MAPE: mecanismos de ausência = MAR 10%; k do kNN = 1; ordem das colunas = *moreCorrelation*; método = Cascata Simplificada.
  - Bias de Correlação: mecanismos de ausência = MAR 20%; k do kNN = 5; ordem das colunas = *noSort*; método = Média.
  - Erro de Similaridade: mecanismos de ausência = MAR 10%; k do kNN = 1; ordem das colunas = *noSort*; método = Cascata Simplificada

- *Forest*
  - MAPE: mecanismos de ausência = MAR 10%; k do kNN = 5; ordem das colunas = *lessMissing*; método = Mediana.
  - Bias de Correlação: mecanismos de ausência = MNAR 10%; k do kNN = 10; ordem das colunas = *lessCorrelation*; método = Cascata Simplificada.
  - Erro de Similaridade: mecanismos de ausência = MAR 10%; k do kNN = 5; ordem das colunas = *moreMissing*; método = Cascata Simplificada.

- *Gamma*
  - MAPE: mecanismos de ausência = MAR 10%; k do kNN = 3; ordem das colunas = *lessCorrelation*; método = Cascata Simplificada.
  - Bias de Correlação: mecanismos de ausência = MAR 20%; k do kNN = 5; ordem das colunas = *moreCorrelation*; método = Cascata Simplificada.
  - Erro de Similaridade: mecanismos de ausência = MAR 10%; k do kNN = 1; ordem das colunas = *moreMissing*; método = Mediana.

- *Glass*
  - MAPE: mecanismos de ausência = MAR 30%; k do kNN = 3; ordem das colunas = *noSort*; método = Mediana.
  - Bias de Correlação: mecanismos de ausência = MNAR 30%; k do kNN = 1; ordem das colunas = *moreMissing*; método = Cascata Simplificada

  - Erro de Similaridade: mecanismos de ausência = MAR 10%; k do kNN = 1; ordem das colunas = *moreMissing*; método = Cascata Simplificada

- *Pima*
  - MAPE: mecanismos de ausência = MAR 10%; k do kNN = 10; ordem das colunas = *lessCorrelation*; método = Cascata Simplificada.
  - Bias de Correlação: mecanismos de ausência = MNAR 10%; k do kNN = 5; ordem das colunas = *moreCorrelation*; método = Cascata Simplificada.
  - Erro de Similaridade: mecanismos de ausência = MAR 10%; k do kNN = 3; ordem das colunas = *moreCorrelation*; método = Mediana.

- *Wine*
  - MAPE: mecanismos de ausência = MAR 10%; k do kNN = 1; ordem das colunas = *lessCorrelation*; método = Cascata Simplificada.
  - Bias de Correlação: mecanismos de ausência = MAR 30%; k do kNN = 5; ordem das colunas = *moreCorrelation*; método = Cascata Simplificada.
  - Erro de Similaridade: mecanismos de ausência = MAR 10%; k do kNN = 1; ordem das colunas = *lessMissing*; método = Cascata Simplificada.

- *Yeast*
  - MAPE: mecanismos de ausência = MAR 10%; k do kNN = 10; ordem das colunas = *lessMissing*; método = Mediana
  - Bias de Correlação: mecanismos de ausência = MCAR 20%; k do kNN = 5; ordem das colunas = *lessCorrelation*; método = Cascata Simplificada
  - Erro de Similaridade: mecanismos de ausência = MAR 10%; k do kNN = 1; ordem das colunas = *lessCorrelation*; método = Mediana

Por meio da análise descritiva dos dados (Tabela V.13), foi possível verificar se as estatísticas (valores mínimos, médias, medianas e máximos) dos dados imputados se mantiveram consistentes com os dados originais.

Nos conjuntos de dados de *Abalone*, *Boston*, *Wine* e *Yeast*, as estatísticas descritivas apresentaram valores semelhantes ou próximos aos dados originais quando utilizamos o Bias de Correlação como métrica de avaliação da imputação. Ao comparar os resultados utilizando o Erro de Similaridade (SIM Erro) como métrica de avaliação, observamos que os dados de *Abalone*, *Gamma* e *Pima* mantiveram distribuições iguais ou próximas aos valores originais. Por fim, apenas os dados

de *Wine* e *Yeast* apresentaram proximidade com os dados originais quando utilizamos o Erro de Imputação (MAPE).

Tabela V.13: Comparativo da análise descritiva entre os dados originais e os imputados

Dados	Estatística	Base Original	Base Imputada		
			MAPE	Bias de Correl.	SIM Erro
Abalone	Mínimo	0,0193	0,0211	0,0193	0,0193
	Média	0,3827	0,3849	0,3822	0,3822
	Mediana	0,3786	0,3792	0,3788	0,3788
	Máximo	1,2391	1,1994	1,2111	1,2111
Boston	Mínimo	17,591	19,0993	19,0993	19,0993
	Média	75,9077	82,0159	81,5893	82,0159
	Mediana	71,1875	77,2956	76,3445	77,2956
	Máximo	127,5303	136,9422	135,4096	136,9422
Gamma	Mínimo	-98,807	-98,807	-93,1721	-98,807
	Média	30,6778	30,6931	30,7315	30,515
	Mediana	28,7102	28,7723	28,7801	28,7502
	Máximo	217,6424	217,6424	217,6424	217,6424
Pima	Mínimo	2,6348	2,6355	2,6355	2,6348
	Média	44,9858	44,6484	44,6397	44,9246
	Mediana	38,3591	37,7331	37,9451	38,5977
	Máximo	179,19	174,69	174,69	179,19
Wine	Mínimo	28,9708	28,9708	28,9708	30,5208
	Média	69,1337	69,1497	69,1497	68,002
	Mediana	63,315	63,2996	63,2996	60,041
	Máximo	148,29	148,29	148,29	145,8862
Forest	Mínimo	5,0333	8,5556	5,4444	10,3556
	Média	93,1714	92,7605	93,0939	91,9425
	Mediana	104,2578	104,2994	103,9356	102,4133
	Máximo	282,6822	282,5378	282,6822	282,0933
Glass	Mínimo	9,7524	9,7846	9,8535	9,7979
	Média	11,2659	11,2291	11,2439	11,2709
	Mediana	11,2892	11,2992	11,292	11,2912
	Máximo	14,2638	14,096	13,5449	14,2127
Yeast	Mínimo	0,1188	0,1188	0,1188	0,1438
	Média	0,3812	0,3804	0,3804	0,3806
	Mediana	0,3675	0,3663	0,3663	0,3685
	Máximo	0,945	0,94	0,94	0,945

Para avaliar se, apesar das diferenças nas estatísticas descritivas, poderíamos considerar que as distribuições dos dados originais e imputados são iguais, foi aplicado o teste não paramétrico de *Wilcoxon*. Na Tabela V.14, observamos que, para os dados de *Boston*, *Gamma*, *Glass* e *Yeast*, os valores imputados apresentaram diferenças estatísticas significativas em relação aos dados originais quando avaliados pelo Bias de Correlação (Bias C.) e pelo Erro de Similaridade (Sim). Dentre os dados mencionados, apenas o conjunto *Boston* teve seus valores imputados identificados como significativamente diferentes dos dados originais pelo teste.

Tabela V.14: Resumo dos testes não paramétrico de Wilcoxon

Dados	Métricas	p-value	Conclusão
Abalone	Mape	0,81	Não há diferença significativa entre original e imputado
	Bias C.	0,1	Não há diferença significativa entre original e imputado
	Sim	0,1	Não há diferença significativa entre original e imputado
Boston	Mape	0	Há uma diferença significativa entre original e imputado
	Bias C.	0	Há uma diferença significativa entre original e imputado
	Sim	0	Há uma diferença significativa entre original e imputado
Gamma	Mape	0,35	Não há diferença significativa entre original e imputado
	Bias C.	0	Há uma diferença significativa entre original e imputado
	Sim	0	Há uma diferença significativa entre original e imputado
Pima	Mape	0,06	Não há diferença significativa entre original e imputado
	Bias C.	0,23	Não há diferença significativa entre original e imputado
	Sim	0,7	Não há diferença significativa entre original e imputado
Wine	Mape	0,38	Não há diferença significativa entre original e imputado
	Bias C.	0,38	Não há diferença significativa entre original e imputado
	Sim	0,92	Não há diferença significativa entre original e imputado
Forest	Mape	0,08	Não há diferença significativa entre original e imputado
	Bias C.	0,16	Não há diferença significativa entre original e imputado
	Sim	0,45	Não há diferença significativa entre original e imputado
Glass	Mape	0,01	Há uma diferença significativa entre original e imputado
	Bias C.	0	Há uma diferença significativa entre original e imputado
	Sim	0,24	Não há diferença significativa entre original e imputado
Yeast	Mape	0	Há uma diferença significativa entre original e imputado
	Bias C.	0	Há uma diferença significativa entre original e imputado
	Sim	0,62	Não há diferença significativa entre original e imputado

Capítulo VI Considerações Finais

Neste capítulo, são apresentadas uma análise retrospectiva sobre a pesquisa realizada, as considerações sobre os resultados encontrados e a sugestão de alguns trabalhos futuros.

VI.1 Análise Retrospectiva

Este estudo busca contribuir para o aprimoramento da metodologia inicialmente desenvolvida por Ferlin [2008]. Entre as inovações sugeridas, as principais são: (i) a análise comparativa envolvendo uma gama mais ampla de algoritmos de agrupamento, sendo eles *Partitioning-based*, *Hierarchical-based*, *Density-based*, e *Model-based*; (ii) a implementação dos mecanismos de ausência NMAR e MAR, além do MCAR já utilizado por Ferlin [2008]; (iii) o desenvolvimento da metodologia Cascata Simplificada, que adapta o modelo original de Ferlin [2008], eliminando a segmentação inicial baseada no padrão de ausência; (iv) a seleção dos conjuntos de dados realizada por meio da classificação das bases mais frequentemente utilizadas em pesquisas acadêmicas mais recentes, acrescentando critérios como a correlação entre suas variáveis e a presença de outliers. As bases foram agrupadas em três categorias: baixa, média e alta, resultando na escolha de oito conjuntos de dados com características distintas; e (v) a formalização da imputação em cascata com o objetivo de torná-la abstrata e independente de implementações específicas de software.

Além disso, o estudo investigou o impacto das características das bases de dados no desempenho do método. Foram analisados aspectos como as topologias das redes, o número de vizinhos utilizados nos algoritmos e os critérios de ordenação, tanto de grupos quanto de atributos, buscando entender como essas variáveis influenciam os resultados obtidos. Essas contribuições visam testar os métodos em diferentes cenários e avaliar sua eficácia sob diversas condições. No total, foram realizadas 335.520 simulações, com uma média de 41.940 instâncias por conjunto de dados, o que destaca a complexidade e a abrangência do experimento realizado para a obtenção dos resultados deste estudo.

VI.2 Visão consolidada dos resultados

A análise dos erros médios e mínimos de imputação nas oito bases de dados revelou que a Imputação em Cascata Simplificada apresentou desempenho superior em várias bases, destacando-se pelo menor erro mínimo, especialmente nas bases *Forest* e *Boston*. No entanto, a imputação

por mediana se mostrou mais eficaz em bases como *Gamma*, especialmente nos mecanismos com maior percentual de valores ausentes, como MCAR 20% e MNAR 30%. Os métodos DBSCAN, Hierárquico, K-modes e SOM apresentaram desempenhos variados, com algumas bases mostrando resultados comparáveis à imputação pela média, dependendo do mecanismo de ausência.

Em relação ao bias de correlação, a Imputação em Cascata Simplificada se destacou por preservar as correlações originais dos dados, com valores de bias de correlação médios e mínimos próximos de zero, indicando pouca alteração nas relações entre os atributos. As bases *Forest*, *Gamma* e *Glass* apresentaram menores biases de correlação nos mecanismos de ausência MAR 10%, MCAR 10% e MNAR 10%. Por outro lado, a imputação univariada, especialmente a média e a mediana, resultou em maiores valores de bias, sugerindo que esses métodos alteram substancialmente as correlações originais.

Quanto aos erros de similaridade, a Imputação em Cascata Simplificada também se mostrou eficaz, especialmente nos mecanismos MCAR 10% e MNAR 10%, apresentando os melhores resultados em termos de similaridade entre os dados imputados e os dados originais. No entanto, a imputação por mediana teve desempenho superior em bases como *Pima* e *Gamma*. Esses resultados indicam que a escolha do método de imputação não deve apenas considerar a precisão na substituição dos valores ausentes, mas também seu impacto na preservação das características fundamentais dos dados, como a similaridade e as correlações entre os atributos.

A análise comparativa entre os erros médios e mínimos em relação aos critérios de ordenação revelou que, enquanto a ordem dos grupos não teve impacto significativo, a ordem dos atributos influenciou de maneira mais pronunciada os erros mínimos, especialmente em algumas bases de dados como *Wine*, *Pima* e *Gamma*. Para esses conjuntos de dados, as variações nos erros mínimos foram mais evidentes, com alguns critérios de ordenação, como *lessCorrelation* e *lessMissing*, apresentando desempenho superior. Nas bases *Boston* e *Abalone*, por exemplo, a ordem dos atributos teve um impacto mais notável, e os critérios *moreCorrelation* e *moreMissing* resultaram em menores erros mínimos. Em contraste, os erros médios apresentaram variações mais suaves, indicando que a ordem dos atributos afeta mais os erros mínimos do que a média global.

Quanto às classes de correlação-outliers, os resultados revelaram padrões distintos de desempenho, dependendo do critério de ordenação e da técnica de imputação. A classe "Baixo C – Baixo O" obteve os melhores resultados com a imputação por mediana e a ordem *lessCorrelation*, enquanto a classe "Médio C – Baixo O" apresentou um desempenho equilibrado entre os métodos e critérios de ordenação. A Imputação em Cascata Simplificada, combinada com a ordem *moreCorrelation*, destacou-se nas classes "Baixo C – Alto O" e "Alto C – Baixo O", enquanto a ordem *noSort* foi mais eficaz nas classes "Baixo C – Médio O" e "Médio C – Médio O". Em algumas classes, como "Médio C – Alto O", a imputação por mediana se destacou, evidenciando a importância de selecionar tanto

o critério de ordenação quanto o método de imputação de forma cuidadosa para obter os melhores resultados.

Quanto ao critério de escolha do valor de k para o algoritmo k-NN, concluímos que a escolha ideal depende da métrica a ser avaliada. Para avaliar o erro de imputação, o valor de $k = 1$ demonstrou o melhor desempenho em 91% dos casos, considerando os diferentes mecanismos de ausência, as classes de correlação-outlier e os critérios de ordenação. Quando o objetivo é avaliar o bias de correlação, o valor de $k = 10$ obteve os melhores resultados em 84% dos casos. Por fim, para a avaliação do erro de similaridade, o valor de $k = 5$ foi o mais eficaz. Esses resultados indicam que a escolha de k deve ser feita com base no tipo de métrica que se deseja otimizar.

Nos conjuntos de dados de Abalone, Boston, Wine e Yeast, o Bias de Correlação mostrou que as estatísticas descritivas dos dados imputados estavam próximas dos valores originais. Usando o Erro de Similaridade (SIM Erro), Abalone, Gamma e Pima mantiveram distribuições semelhantes às originais, enquanto Wine e Yeast apresentaram proximidade no Erro de Imputação (MAPE). O teste de Wilcoxon foi aplicado para avaliar a equivalência das distribuições. Os resultados indicaram diferenças significativas nos dados de Boston, Gamma, Glass e Yeast, mas apenas o conjunto Boston teve os valores imputados identificados como significativamente diferentes dos originais.

VI.3 Trabalhos Futuros

Neste trabalho, foi realizada uma análise detalhada da técnica de imputação em cascata, com foco na sua aplicação em bases compostas por atributos contínuos. No entanto, uma extensão interessante dessa abordagem seria aplicar a mesma técnica a bases compostas por atributos categóricos e bases mistas, que combinam tanto atributos contínuos quanto categóricos. A razão para isso é que bases com atributos categóricos apresentam desafios diferentes em termos de imputação, especialmente em relação à escolha de métodos adequados para lidar com variáveis discretas. Estudar como a imputação em cascata se comporta nessas bases pode revelar *insights* sobre a adaptabilidade e a eficácia da técnica em cenários mais complexos.

Outro ponto relevante seria a avaliação do desempenho da imputação em cascata em grandes bases de dados, uma vez que, neste trabalho, não foi possível realizar a imputação para bases de tamanho significativamente grande, devido à limitação de tempo e recursos computacionais. Uma investigação mais profunda sobre se o esforço computacional para a imputação é justificável em cenários de grandes volumes de dados poderia fornecer informações valiosas. É possível que, para bases muito grandes, a estratégia mais eficiente seja simplesmente excluir as tuplas nulas, especialmente quando o custo computacional de imputar os valores ausentes seja elevado. Estudar essa questão ajudaria a determinar em que cenários a imputação em cascata realmente oferece benefícios em termos de qualidade dos dados e desempenho, em comparação com abordagens mais

simples, como a exclusão de dados ausentes.

Uma área que poderia ser explorada em futuros estudos é a avaliação de diferentes métodos de cálculo de distâncias, além da Euclidiana e *Gower*, que foram utilizadas neste trabalho. Distâncias como *Manhattan*, *Mahalanobis*, entre outras, têm características que podem ser mais apropriadas para determinados tipos de dados, especialmente em bases mistas ou em casos onde as variáveis possuem escalas diferentes. Testar essas diferentes métricas poderia permitir uma análise mais robusta e precisa do desempenho da imputação em cascata, pois a escolha da métrica de distância pode impactar diretamente a qualidade da imputação, especialmente quando se trata de preservar as relações entre os dados ao longo das etapas de imputação. Essa investigação pode oferecer novas perspectivas sobre como otimizar a técnica de imputação para uma variedade de cenários de dados.

Outro aspecto importante seria determinar o valor ideal de k para o algoritmo k-NN (k-vizinhos mais próximos). Atualmente, o valor de k é frequentemente determinado de maneira empírica, mas uma abordagem interessante seria utilizar heurísticas, como $k = \sqrt{N}$ (onde N é o número de amostras da base) ou métricas de análise de agrupamento, para otimizar o desempenho do k-NN para cada atributo individualmente. Essa abordagem poderia melhorar a precisão da imputação ao considerar a estrutura subjacente dos dados de forma mais dinâmica, ajustando o valor de k conforme as características específicas de cada variável.

Uma análise da acurácia dos valores imputados poderia ser realizada por meio da reclassificação dos casos, especialmente em bases que possuam um atributo-classe. Medir a acurácia das imputações pode ajudar a avaliar a eficácia da técnica de imputação em cascata não apenas com base em medidas estatísticas, mas também com uma métrica mais próxima da aplicação prática, verificando se as imputações realmente melhoram ou preservam a classificação dos casos.

Outra linha de pesquisa poderia ser a troca dos algoritmos utilizados nos pipelines de imputação. Embora este trabalho tenha se concentrado em determinados métodos, a substituição ou combinação de algoritmos para tratar os dados faltantes pode gerar melhorias significativas no desempenho da imputação. Explorar como diferentes algoritmos, como os baseados em redes neurais, métodos supervisionados e semi-supervisionados ou métodos de aprendizado profundo, podem ser integrados aos processos de imputação pode resultar em abordagens mais eficientes, especialmente em contextos onde os dados são complexos e de alta dimensionalidade.

Finalmente, uma direção promissora seria a aplicação da metodologia de imputação em cascata em áreas além do domínio clássico de imputação de dados. Áreas como processamento de sinais, completação ou diagonalização de matrizes, citometria e outras disciplinas científicas podem se beneficiar do uso da imputação em cascata. Essas áreas, embora frequentemente fora do escopo das técnicas tradicionais de imputação, podem apresentar desafios únicos que exigem adaptações da técnica, abrindo novas possibilidades para o uso dessa abordagem em problemas interdisciplinares

complexos. Investigar como a imputação em cascata pode ser aplicada nessas áreas poderia expandir significativamente o alcance da técnica e fornecer *insights* inovadores sobre o tratamento de dados ausentes em contextos especializados.

Referências

Aha, W., Kibler, D., and Albert, M. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

Ahmad, A. and Khan, S. S. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7:31883–31902, 2019.

Aroyo, L., Lease, M., Paritosh, P. K., and Schaekermann, M. Data excellence for AI: why should you care. *CoRR*, abs/2111.10391, 2021.

Assunção, F. Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos. Master's thesis, Universidade de Sao Paulo, 2012.

Bano, S. and Khan, N. A survey of data clustering methods. *International Journal of Advanced Science and Technology*, 113, 2018.

Brown, M. and Kros, J. Data mining and the impact of missing data. *Industrial Management and Data Systems*, 103:611–621, 2003.

Bussab, W. d. O. and Moretin, P. A. *Estatística básica*. Saraiva, São Paulo, 9.ed. edition, 2017.

Cartwright, M., Shepperd, M., and Song, Q. Dealing with missing software project data. *Software Metrics, IEEE International Symposium on*, 0:154, 2003.

Castañeda, R., Ferlin, C., Goldschmidt, R., Soares, J., Carvalho, L., and Choren, R. Aprimorando processos de imputação multivariada de dados com workflows. *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 238–252, 2008.

Ciaburro, G. and Venkateswaran, B. *Neural Networks with R : smart models using CNN, RNN, deep learning, and artificial intelligence principles*. Packt, Birmingham, England, 1st edition edition, 2017 - 2017.

Colin-Chevalier, R., Dutheil, F., Cambier, S., Dewavrin, S., Cornet, T., Baker, J. S., and Pereira, B. Methodological issues in analyzing real-world longitudinal occupational health data: A useful guide to approaching the topic. *International Journal of Environmental Research and Public Health*, 19(12), 2022.

Coneglian, C., Gonçalez, P., and Santarem Segundo, J. O profissional da informação na era do big data. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 22:128, 2017.

Coombes, C., Abrams, Z., Coombes, K., and Brock, G. Simulation-derived best practices for clustering clinical data. *Journal of biomedical informatics*, 118:103788, 2021.

Deza, M. M. and Deza, E. *Encyclopedia of distances*. Springer, Berlin, 2009.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A., Foufou, S., and Bouras, A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *Emerging Topics in Computing, IEEE Transactions on*, 2:267–279, 2014.

Fausett, L. *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall, Inc., USA, 1994.

Ferlin, C. *Imputação em cascata: uma abordagem para imputação multivariada de dados*. PhD thesis, Universidade Federal do Rio de Janeiro - COPPE, Rio de Janeiro, RJ, 2008.

Fontes, N., Almeida, J., and Silva, G. d. Utilização do big data para obter vantagens competitivas. *Revista Científica On-line Tecnologia Gestão Humanismo*, 6:111–120, 2016.

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2006.

Gonçalves, L. M. Imputação hot-deck: Uma revisão sistemática da literatura. Master's thesis, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), 2021.

Graham, J W, . D. S. I. Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, 78 1:119–128, 1993.

Graham, J. *Missing data: Analysis and design*. Springer, 2012.

Graham, J. W., Cumsille, P. E., and Shevock, A. E. *Methods for Handling Missing Data*, chapter 4. John Wiley Sons, Ltd, 2012.

Jain, A. Data clustering: 50 years beyond k-means. *King-Sun Fu prize lecture at the 19th international conference on pattern recognition*, pages 3–4, 2008.

Jarrahi, M. H., Memariani, A., and Guha, S. The principles of data-centric ai (dcai). *Communications of the ACM*, 2022.

Jonsson, P. and Wohlin, C. An evaluation of k-nearest neighbour imputation using likert data. *Proceedings - International Software Metrics Symposium*, pages 108 – 118, 2004.

Kar, A., Mishra, A., and Mohanty, S. K. An efficient entropy based dissimilarity measure to cluster categorical data. *Engineering Applications of Artificial Intelligence*, 119:105795, 2023.

Larose, D. and Larose, C. *Data Mining and Predictive Analytics*. Wiley Series on Methods and Applications in Data Mining, 2015.

Little, R. and Rubin, D. *Statistical Analysis with Missing Data*. Wiley, 2019.

Magnani, M. Techniques for dealing with missing data in knowledge discovery tasks. *Disponível em: https://api.semanticscholar.org/CorpusID:15408337*, 2004.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Mitchell, T. M. *Machine learning*, volume 1. McGraw-hill New York, 1997.

Montiel, J., Read, J., Bifet, A., and Abdessalem, T. Scalable model-based cascaded imputation of missing data. In Phung, D. Q., Tseng, V. S., Webb, G. I., Ho, B., Ganji, M., and Rashidi, L., editors, *PAKDD (3)*, volume 10939 of *Lecture Notes in Computer Science*, pages 64–76. Springer, 2018.

Musolf, A., Holzinger, E., Malley, J., and Bailey-Wilson, J. What makes a good prediction? feature importance and beginning to open the black box of machine learning in genetics. *Human Genetics*, 141, 2022.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

Raghunathan, T. What do we do with missing data? some options for analysis of incomplete data. *Annual review of public health*, 25:99–117, 2004.

Ran, X., Xi, Y., Lu, Y., Wang, X., and Lu, Z. Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, 56:1–46, 2022.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Rubin, D. B. An overview of multiple imputation. *Journal of the American Statistical Association*, page 79–84, 1988.

Schafer, J. and Graham, J. Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177, 2002.

Soares, J. *Pré-Processamento em Mineração de Dados: Um Estudo Comparativo em Complementação*. PhD thesis, COPPE/UFRJ - Engenharia de Sistemas e Computação, Rio de Janeiro, RJ, 2007.

van Buuren, S. *Flexible Imputation of Missing Data*. Chapman Hall/CRC Press, Boca Raton, FL, 2012.

Veroneze, R. *Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla*. PhD thesis, UNICAMP - Faculdade de Engenharia Elétrica e de Computação, Campinas, SP, 2011.

Wang, Y., Li, B., Luo, R., Chen, Y., Xu, N., and Yang, H. Energy efficient neural networks for big data analytics. *2014 Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1–2, 2014.

Wayman, J. C. Multiple imputation for missing data: What is it and how can i use it. *Annual Meeting of the American Educational Research Association, Chicago, IL*, 2:16, 2003.

Willmott, C. J. On the validation of models. *Physical Geography*, 2(2):184–194, 1981.

Xu, D. and Tian, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2, 2015.

Yim, O. and Ack Baraly, K. Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology*, 11:8–21, 2015.

Apêndice A : Configurações utilizadas como Critério de Ordenações

Tabela A.1: Descritivo das configurações utilizadas como critério de ordenações

Ordem do Grupos	Ordem dos Atributos	Configuração
fieldLessMissing	lessCorrelation	A - A
	lessMissing	A - B
	moreCorrelation	A - C
	moreMissing	A - D
	noSort	A - E
fieldMoreMissing	lessCorrelation	B - A
	lessMissing	B - B
	moreCorrelation	B - C
	moreMissing	B - D
	noSort	B - E
fieldPerTupleLessMissing	lessCorrelation	C - A
	lessMissing	C - B
	moreCorrelation	C - C
	moreMissing	C - D
	noSort	C - E
fieldPerTupleMoreMissing	lessCorrelation	D - A
	lessMissing	D - B
	moreCorrelation	D - C
	moreMissing	D - D
	noSort	D - E
noSort	lessCorrelation	E - A
	lessMissing	E - B
	moreCorrelation	E - C
	moreMissing	E - D
	noSort	E - E
random	lessCorrelation	F - A
	lessMissing	F - B
	moreCorrelation	F - C
	moreMissing	F - D
	noSort	F - E
tupleLessMissing	lessCorrelation	G - A
	lessMissing	G - B
	moreCorrelation	G - C
	moreMissing	G - D
	noSort	G - E
tupleMoreMissing	lessCorrelation	H - A
	lessMissing	H - B
	moreCorrelation	H - C
	moreMissing	H - D
	noSort	H - E

Apêndice B : Comparação das Configurações dos Critério de Ordenações e as Classes Correlação-Outlier

Os códigos referentes às configurações das classes de correlação e *outliers* foram utilizados na Tabela B.1 a fim de tornar a apresentação mais compacta. As conclusões sobre o comparativo do erro médio por configurações dos critérios de ordenações e classe correlação-outlier se encontra na Seção V.4.3.

Tabela B.1: Comparativo do erro médio por configurações dos critérios de ordenações e classe correlação-outlier (base e^{12})

Configuração	0 - 0	0 - 1	0 - 2	1 - 0	1 - 1	2 - 2	2 - 0	2 - 1
A - A	4252,96	227,93	0,51	0	16,19	43,9	0,0102	266,5
A - B	4302,83	243,82	0,51	0	16,02	43,85	0,0102	256,68
A - C	4302,85	245,24	0,49	0	15,99	44,9	0,0097	259,8
A - D	4277,47	236,92	0,51	0	16,09	43,91	0,0098	260,21
A - E	4290,57	227,58	0,53	0	15,92	44,89	0,0099	261,15
B - A	4252,96	227,93	0,51	0	16,19	43,9	0,0102	266,5
B - B	4302,83	243,82	0,51	0	16,02	43,85	0,0102	256,68
B - C	4302,85	245,24	0,49	0	15,99	44,9	0,0097	259,8
B - D	4277,47	236,92	0,51	0	16,09	43,91	0,0098	260,21
B - E	4290,57	227,58	0,53	0	15,92	44,89	0,0099	261,15
C - A	4252,96	227,93	0,51	0	16,19	43,9	0,0102	266,5
C - B	4302,83	243,82	0,51	0	16,02	43,85	0,0102	256,68
C - C	4302,85	245,24	0,49	0	15,99	44,9	0,0097	259,8
C - D	4277,47	236,92	0,51	0	16,09	43,91	0,0098	260,21
C - E	4290,57	227,58	0,53	0	15,92	44,89	0,0099	261,15
D - A	4252,96	227,93	0,51	0	16,19	43,9	0,0102	266,5
D - B	4302,83	243,82	0,51	0	16,02	43,85	0,0102	256,68
D - C	4302,85	245,24	0,49	0	15,99	44,9	0,0097	259,8
D - D	4277,47	236,92	0,51	0	16,09	43,91	0,0098	260,21
D - E	4290,57	227,58	0,53	0	15,92	44,89	0,0099	261,15
E - A	4252,96	227,93	0,51	0	16,19	43,9	0,0102	266,5
E - B	4302,83	243,82	0,51	0	16,02	43,85	0,0102	256,68
E - C	4302,85	245,24	0,49	0	15,99	44,9	0,0097	259,8
E - D	4277,47	236,92	0,51	0	16,09	43,91	0,0098	260,21
E - E	4290,57	227,58	0,53	0	15,92	44,89	0,0099	261,15
F - A	4252,96	227,93	0,51	0	16,19	43,9	0,0102	266,5
F - B	4302,83	243,82	0,51	0	16,02	43,85	0,0102	256,68
F - C	4302,85	245,24	0,49	0	15,99	44,9	0,0097	259,8
F - D	4277,47	236,92	0,51	0	16,09	43,91	0,0098	260,21
F - E	4290,57	227,58	0,53	0	15,92	44,89	0,0099	261,15
G - A	4252,96	227,93	0,51	0	16,19	43,9	0,0102	266,5
G - B	4302,83	243,82	0,51	0	16,02	43,85	0,0102	256,68
G - C	4302,85	245,24	0,49	0	15,99	44,9	0,0097	259,8
G - D	4277,47	236,92	0,51	0	16,09	43,91	0,0098	260,21
G - E	4290,57	227,58	0,53	0	15,92	44,89	0,0099	261,15
H - A	4252,96	227,93	0,51	0	16,19	43,9	0,0102	266,5
H - B	4302,83	243,82	0,51	0	16,02	43,85	0,0102	256,68
H - C	4302,85	245,24	0,49	0	15,99	44,9	0,0097	259,8
H - D	4277,47	236,92	0,51	0	16,09	43,91	0,0098	260,21
H - E	4290,57	227,58	0,53	0	15,92	44,89	0,0099	261,15